Proceedings of the Institute of Acoustics

# A Study of Vocal Tract Shapes Using Magnetic Resonance Imaging and Acoustic Reflectance Techniques

J. W. Devaney & C. C. Goodyear

Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK

## 1. INTRODUCTION

A major difficulty in performing speech synthesis using an articulatory model is the lack of sufficient data on the range of vocal tract shapes. The vocal tract is usually represented as a concatenation of short uniform tubes, the areas of which are referred to as the vocal tract area function. In order to produce synthetic speech, a sequence of these area functions is required, which when used in the synthesiser, will mimic a given natural utterance.

A commonly used technique for obtaining a sequence of area functions employs a very large code book, randomly populated with vocal tract shapes. The task of searching such a code book is computationally expensive, though this cost can be somewhat reduced using a neural net mapping method [1]. For copy synthesis of a single speaker however, a much more efficient code book could be generated if area functions close to those used by the speaker were known.

In some earlier work [2], magnetic resonance imaging was used to determine the vocal tract area functions of a male speaker producing 5 steady state vowel sounds. The data was used in an acoustic tube simulation, and the resulting acoustic spectra were compared with those of natural speech from the same speaker. For all but one of the vowels good agreement, typically within 120Hz, was obtained for the first three formant frequencies.

Magnetic resonance imaging is however costly and time consuming. Moreover, a far larger number of shapes than could be obtained from imaging is needed for the task of acoustic to geometric mapping. For this reason, we investigated an acoustic technique, first described by Sondhi and Gopinath [3]. Preliminary results [4] showed that the basic acoustic technique, when used in conjunction with further optimisation, provided a low cost method of extracting both rubber-model and real vocal tract area functions.

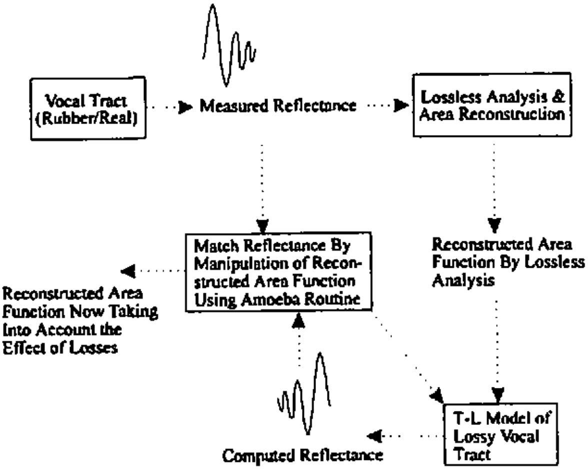
Further experiments have allowed us to study the changing vocal tract shape during diphthongs. A parametric model of the vocal tract was employed to simulate the mechanism of the vocal tract during these transitions.

## 2. THE ACOUSTIC REFLECTANCE TECHNIQUE

In this method, an acoustic volume-velocity impulse is directed towards the lips of a chosen speaker while the subject articulates a particular vowel with closed glottis. The reflected pressure wave is then recorded. By analysis of this reflected wave, the area function of the vocal tract may be reconstructed. Details of the analysis technique may be found in [3] and [4].

The basic analysis technique [3] assumes that the wave propagation within the vocal tract is planar, and that the vocal tract is lossless. However in a real acoustic tube, losses occur due to viscous friction, heat conduction, and wall vibration. Sondhi and Resnick [5] were able to model two cases in which losses were included : case 1, for a vocal tract with yielding walls and no viscous loss and case 2, for a vocal tract with viscous loss and hard walls. To overcome the effect of all three sources of loss, we have used the basic no-loss analysis to give a starting point for an optimisation procedure. This is described in figure 2.1.

*Figure 2.1 : Optimisation Schematic to Overcome Losses*

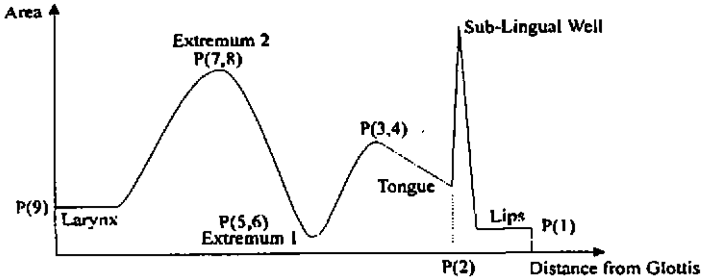

Vocal Tract (Rubber/Real)
Measured Reflectance
Lossless Analysis & Area Reconstruction
Match Reflectance By Manipulation of Reconstructed Area Function Using Amoeba Routine
Reconstructed Area Function By Lossless Analysis
Reconstructed Area Function Now Taking Into Account the Effect of Losses
T-L Model of Lossy Vocal Tract
Computed Reflectance

The diagram shows the measured reflected pressure wave from the vocal tract first analysed by the technique described in [3]. This yields a reconstructed area function, but one which cannot be identical to the original as losses have not been considered. The reconstructed area function is therefore placed back into a lossy transmission line model of the vocal tract [6] whose impulse reflectance is computed. A recursive optimisation procedure is used to match the computed reflectance to the measured reflectance by manipulation of the area function. Once the two reflectances are sufficiently well matched, the optimised area function must be close to the original area function of the vocal tract, assuming that the transmission line model is a sufficiently accurate representation of the vocal tract being analysed.

## 3. PARAMETRIC MODELLING OF THE VOCAL TRACT

While it is possible to specify each area independently, this would ignore anatomical constraints A much better approach is use some form of parametric model of the vocal tract. The parameters of this model may be used to generate an area profile, which when sampled at a constant spacing, provides areas for driving a synthesiser. The parametric model [7] utilised in this paper uses nine parameters, and has the form shown in figure 3.1.

*Figure 3.1 : The Nine Parameter Vocal Tract Model*

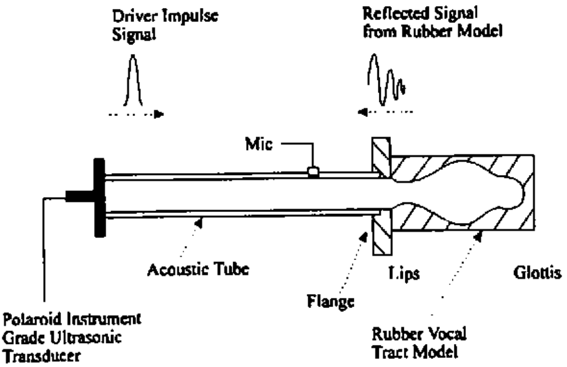


The parameters are : the lip area, the cross-sectional area at the tongue tip, the position of the tongue blade rear, the cross-sectional area at the blade rear, the position of extremum 1, cross-sectional area at extremum 1, position of extremum 2, cross-sectional area at extremum 2, and the area at the larynx. Extrema 1 and 2 refer to positions along the vocal tract profile where the gradient is zero. The length of the tongue blade is fixed at 2cm, and both the larynx section length and lip length are fixed to 0.875cm. The position of the sub-lingual well is dependent upon the position of the tongue tip, and the area at this point is constrained to a maximum of 8cm$^2$.

## 4. ACOUSTIC EXPERIMENTS

Figure 4.1 shows the apparatus used for the acoustic experiment. An ultrasonic transducer was used to generate an impulse in volume-velocity. This impulse was not ideal, and our reflectance measurements were corrected in the frequency-domain to compensate for its non-ideal spectrum up to our band limit of 5kHz. The corrected reflectance was sampled at a rate of 10kHz. With velocity of sound $c$ = 35000 cm/sec, this implies an area function resolution of 10 concatenated sections, each 1.75cm long.

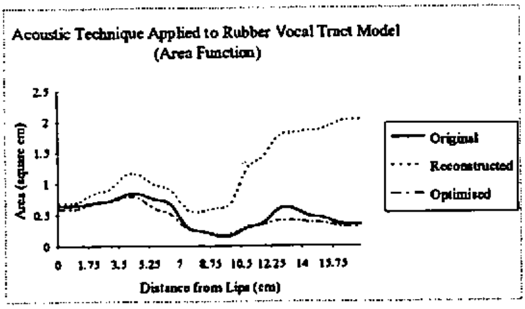
*Figure 4.1 : Acoustic Apparatus*


Driver Impulse Signal
Reflected Signal from Rubber Model
Mic
Acoustic Tube
Lips
Glottis
Flange
Polaroid Instrument Grade Ultrasonic Transducer
Rubber Vocal Tract Model

**4.1 Evaluation of Acoustic Reflectance Method.**

To evaluate the acoustic analysis and optimisation procedure, initial experiments were performed using rubber vocal tract models. The rubber models comprised three shapes, and the use of two materials. The wall yielding losses of each material were initially set to values of $r_S$=1000Ω and $l_S$=2H. These values were then adapted using a recursive procedure to improve the geometric match. The adapted losses (different for each material) were kept constant for each of the shapes. Results for one of the rubber models are shown in figure 4.2. Further results using rubber vocal tract models may be found in [4].

*Figure 4.2 : Acoustic Technique Applied to Rubber Vocal Tract Model*



It was found for both materials, and all shapes, that the effect of losses on the computed shapes was quite substantial. However, the optimisation method described in section 2 was found to overcome this effect in all cases.

### 4.2 Analysis of a Real Vocal Tract for Sustained Vowels

The investigation was continued using a real vocal tract. In this case, the subject silently articulated a particular vowel while the acoustic reflectance was captured. The reflectance was used to compute a 'real' vocal tract shape using the acoustic technique. To estimate the area at the lips, the lip shape was assumed to be an ellipse, and its area was calculated by direct measurements of the major and minor axis. The wall yielding loss parameters for the optimisation procedure were fixed at $r_S$=1000Ω and $l_S$=2H.

Six vowels were studied, /IY/, /ER/, /AR/, /AE/, /OO/ and /OR/, taken from the words *"heat"*, *"heard"*, *"hard"*, *"had"*, *"who'd"*, and *"hawk"* respectively. In each case, the shape was used in a Kelly Lochbaum model with losses, as employed by Geenwood [2]. The loss factor in each section is such that it crudely approximates the frequency dependent losses present in the vocal tract. The formant frequencies were identified from the models impulse response, and are listed in table 4.1.

*Table 4.1 : Formant Frequencies for Natural and Synthetic Speech*

| Vowel | Formant | Natural Formants (Hz) | Synthetic Formants (Hz) | Error (Hz) |
|---|---|---|---|---|
| /IY/ | 1 | 291 | 448 | -157 |
| | 2 | 2138 | 1775 | 363 |
| | 3 | 2924 | 2447 | 477 |
| /ER/ | 1 | 577 | 528 | 49 |
| | 2 | 1403 | 1377 | 26 |
| | 3 | 2671 | 2552 | 119 |
| /AR/ | 1 | 767 | 680 | 87 |
| | 2 | 980 | 1228 | 248 |
| | 3 | 2747 | 2391 | 356 |
| /AE/ | 1 | 775 | 598 | 177 |
| | 2 | 1700 | 1575 | 125 |
| | 3 | 2408 | 2373 | 35 |
| /OO/ | 1 | 265 | 528 | -263 |
| | 2 | 741 | 965 | -224 |
| | 3 | 2012 | 2239 | -227 |
| /OR/ | 1 | 488 | 655 | -177 |
| | 2 | 733 | 739 | -6 |
| | 3 | 2170 | 1810 | 360 |

The overall rms. error between natural and synthetic formants is 233Hz. A possible source of error in the analysis technique was thought to be the modelling of vocal tract losses during optimisation. To improve the match with the subjects natural spectrum, a second stage of optimisation was therefore applied to the reconstructed area function which incorporated both geometric and spectral penalties. The spectral penalty employed was the mean square difference between the first three synthetic and natural formants. The geometric penalty was taken to be the mean square difference between the adapted area function and the original starting point area function. The geometric penalty was weighted by an empirically chosen value, which for the results reported here was 0.1.

Figures 4.3(a) to 4.5(a) show the resulting vocal tract shapes for the vowels /AR/, /IY/, and /AE/, where the area functions are compared with those obtained by magnetic resonance imaging for the same vowel.

Fig 4.3(a) : Comparison of Acoustic & Magnetic Resonance Reconstructed Area Functions

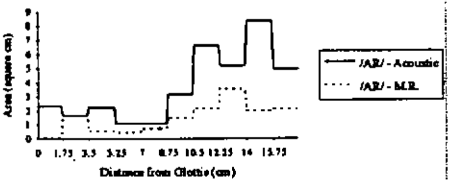


Fig 4.3(b) : Comparison of Impulse Spectra Generated by Acoustic & M.R. Area Functions to Natural Spectra

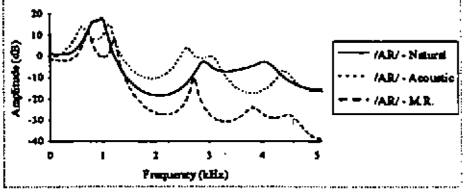


Fig 4.4(a) : Comparison of Acoustic & Magnetic Resonance Reconstructed Area Functions

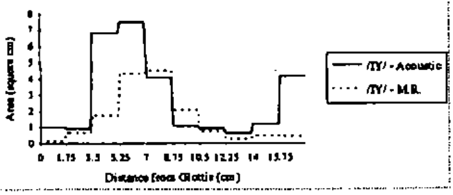


Fig 4.4(b) : Comparison of Impulse Spectra Generated by Acoustic & M.R. Area Functions to Natural Spectra

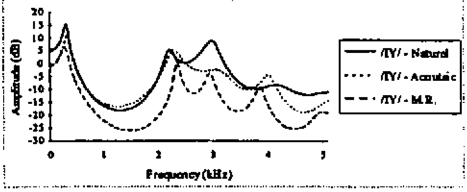


Fig 4.5(a) : Comparison of Acoustic & Magnetic Resonance Reconstructed Area Functions

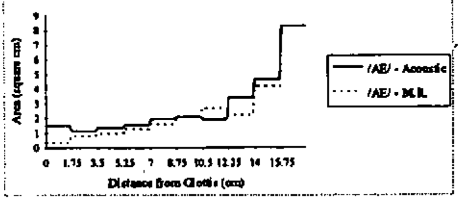
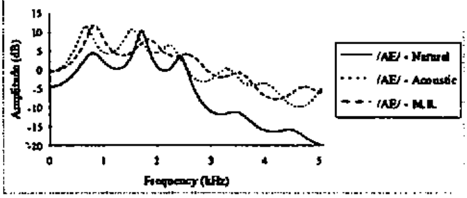


Fig 4.5(b) : Comparison of Impulse Spectra Generated by Acoustic & M.R. Area Functions to Natural Spectra



Figures 4.3(b) to 4.5(b) show a comparison of the synthetic spectra to the natural spectra of the same speaker. The acoustic-tube model of the vocal tract assumes plane wave propagation, therefore the spectral comparison is only valid for frequencies up to about 3.5kHz. Although this upper frequency limit may be reduced somewhat due to the finite section length of the model.

### 4.3 Analysis of a Real Vocal Tract During Gestures

In this case, a sequence of six acoustic reflectances was captured while the subject silently articulated a particular diphthong. The vocal tract shape was computed using the acoustic technique. With view to obtaining smoothly evolving shapes, the parametric model was applied to each of the reconstructed area functions during the gesture. This was achieved by optimising the nine parameters of the model to fit the reconstructed area function whilst targeting on the natural formant positions at a corresponding instant in the gesture. The optimisation procedure used the same cost penalty as that described for the sustained vowel case.

The natural formant track was obtained by recording the utterance, and using a cepstral smoothing and peak picking algorithm to determine the formants. Ten diphthongs were studied. Typical results obtained for the transition of /IY/ to /ER/ located in the word "h*ear*", and /OR/ to /I/ located in the word *"foil"* are shown in figures 4.6 and 4.7.

*Figure 4.6 : Vocal Tract Gesture /IY/ - /ER/, Area Function & Impulse Spectra*

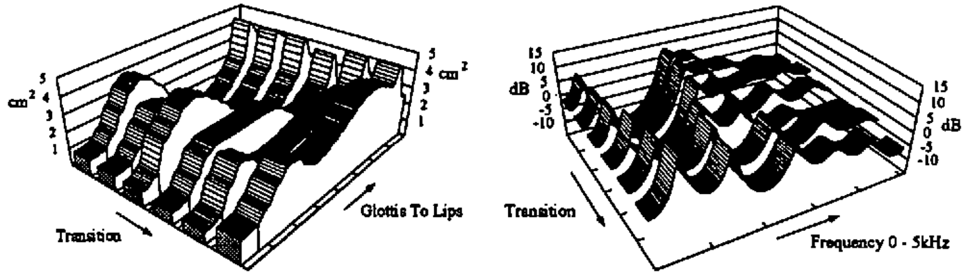
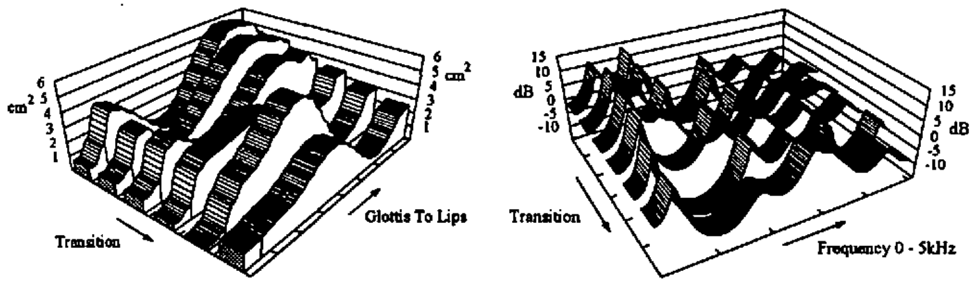


*Figure 4.7 : Vocal Tract Gesture /OR/ - /I/, Area Function & Impulse Spectra*



## 5. CONCLUSIONS

The similarity between the area functions derived from the acoustic and MR methods is encouraging, although it is not known how large were any real differences in the speaker's articulations of the same vowel under the very different experimental conditions. The changes to the area functions after optimisation to the natural spectra were generally small, due to the geometric penalty. However, these changes were sufficient to reduce the overall rms error between the natural and synthetic formants from 233Hz to 88Hz for the six vowels.

The investigation into the changing vocal tract shape during diphthongs, and the use of parametric modelling, has provided smooth transitions in the vocal tract area function. These data when used to synthesise speech were found to generate formant tracks similar to the natural diphthongs.

## 6. ACKNOWLEDGEMENTS

The authors are grateful to British Telecom for their financial assistance under the CASE award scheme administered by SERC.

## 7. REFERENCES

[1] RAHIM M G, GOODYEAR C C, KLEIJN W B, SCHROETER J, & SONDHI M M, 'On the Use of Neural Networks in Articulatory Speech Synthesis', J. Acoust. Soc. Amer, Vol. 93, pp 1109-1121, 1993.

[2] GREENWOOD A R, GOODYEAR C C, & MARTIN P A, 'Measurements of Vocal Tract Shapes using Magnetic Resonance Imaging', IEE. Proceedings-1, Vol. 139, No. 6, 1992.

[3] SONDHI M M & GOPINATH B, 'Determination of Vocal-Tract Shape from Impulse Response at the Lips', J. Acoust. Soc. Amer., Vol. 49, pp 1867-1873, 1971.

[4] DEVANEY J W & GOODYEAR C C, 'A Comparison of Acoustic and Magnetic Resonance Imaging Techniques in the Estimation of Vocal Tract Area Functions', IEEE. Symposium, ISSIPNN, Vol. 2, pp 575-578, Hong Kong, 1994.

[5] SONDHI M M & RESNICK J R, 'The Inverse Problem for the Vocal Tract: Numerical Methods, Acoustical Experiments, and Speech Synthesis', J. Acoust. Soc. Amer, Vol. 73(3), pp 985-1002, 1983.

[6] FLANAGAN J L, 'Speech Analysis, Synthesis and Perception', 2nd Edition, Springer - Verlag, New York, 1972.

[7] GREENWOOD A R & GOODYEAR C C, 'Articulatory Speech Synthesis Using a Parametric Model and a Polynomial Mapping Technique', IEEE. Symposium, ISSIPNN, Vol. 2, pp 595-598, Hong Kong, 1994.