A New Algorithm for Vocal Tract Shape Extraction from Singer's Waveforms

Rebecca Vos\(^1\), Jamie Andrea Shyla Angus\(^2\), and Brad H Story\(^3\)

\(^1\) School of Computing, Science and Engineering, University of Salford, Salford, Greater Manchester, M5 4WT, United Kingdom  
    rebecca_vos@york.ac.uk

\(^2\) School of Computing, Science and Engineering, University of Salford, Salford, Greater Manchester, M5 4WT, United Kingdom  
    j.a.s.angus@salford.ac.uk

\(^3\) Speech Acoustics Laboratory, Department of Speech and Hearing Sciences, University of Arizona, Tucson, Arizona 85721  
    bstory@email.arizona.edu

ABSTRACT

This paper presents a new algorithm for extracting vocal tract shape from, speech or singing. Based on acoustic sensitivity functions it removes the ambiguity that conventional methods suffer from. We describe acoustic sensitivity functions and how we extract the necessary formant frequencies from the acoustic waveform. Results are presented for a variety of singers both male and female singing a variety of vowels and notes. The results are good and the system not only has applications in voice training but could also be used for control of games or music synthesis.

1. INTRODUCTION

Singers often have to produce a sound level that cuts through the surrounding musical arrangement, and a variety of singing styles such as "Belting", "Bel Canto", etc. have been developed to allow a singer to do this without damaging their voice.

Since it is known that one of the principle determining factors on the acoustic output of the vocal system is the vocal tract, being able to see a real-time (or close to real-time) representation of the shape of the vocal tract...
would allow the user to consciously modify its shape, improving their communication skills, or achieving a “better” tone when singing. This could involve more accurate pitching, or simply a clearer or more pleasing timbre.

In the past, MRI machines have been used to scan the patient’s head, and thus extract the shape of their vocal tract. This is a highly accurate process, and in an ideal world would be the primary method of determining the shape of a vocal tract, however in reality it is time-consuming and expensive, and must be done in a hospital, meaning that it cannot be done in real-time or even close to it, and is not accessible for the average speech therapist or vocal coach.

With modern advances in computing, it is now possible to create a virtual model of a vocal tract by concatenating short cylinders, and there are several programs available that allow the user to alter the shape of a virtual vocal tract, and hear the output from it, and even see the spectrum of the acoustic output.

However, doing this in reverse, i.e. taking an acoustic signal and then deducing the shape of the vocal tract that must have produced it, is more complicated. One of the chief complications is the “ventriloquist effect” where the same sound can be produced from several different vocal tract shapes. This means there is a many-to-one relationship between the acoustic tube area profile and the output spectral shape. This would seem to make the idea of an acoustic waveform to area model program impossible.

In 2006 Brad H. Story [1] developed a method that used sensitivity functions to reduce these errors. However his method, though effective, required that the formant, or resonance, frequencies be measured and entered manually.

This paper takes this work further and presents a system that first extracts the requisite formant frequencies from the singer's acoustic waveform and then uses Story's method to extract the probable vocal tract shape.

Firstly we will describe the ambiguity problem and then discuss how the use of sensitivity functions can reduce this problem.

It will present both the details of the new algorithm and the results of testing it on a variety of male and female singers, both trained and untrained, singing a variety of vowels at several different pitches.

2. WHAT IS THE PROBLEM?

The main problem with deriving vocal tract shape from the speech waveform is that it is a many-to-one mapping in that a multiplicity of vocal tract configurations can produce similar spectra. This is well known and has been demonstrated by many researchers. For example, Mermelstein, [2]; Schroeder, [3]; Wakita, [4]; Atal, Chang, Mathews, and Tukey, [5]; Sondhi and Resnick, [6]; Milenkovic, [7]; and Panchapagesan, [8].

\[
\begin{align*}
F_1 &= 275 \ Hz, \ F_2 &= 2132 \ Hz, \ F_3 &= 2998 \ Hz, \ F_4 &= 4412 \ Hz \\
I_2 &= 8 \ cm, \ I_1 &= 6 \ cm, \ A_2 &= 8 \ cm^2, \ A_1 &= 1 \ cm^2
\end{align*}
\]

This ambiguity is often exploited by a ventriloquist to make equivalent sounds without changing their mouth opening.
However, as also discussed in Fant [9], in many cases the different vocal tract shapes required to be ambiguous are physiologically difficult, if not impossible for a human to produce. In addition an empirical study by Qin and Carreira-Perpinan [10] concluded that although nonuniqueness occurred for some sounds in real speech it was the exception rather than the rule. Nonetheless, the possibility of ambiguity is still an issue with any simplistic approach to deriving vocal tract shape from the speech waveform and thus a more sophisticated approach is required. One such approach is the use of acoustic sensitivity functions.

3. **WHY MIGHT ACOUSTIC SENSITIVITY FUNCTIONS WORK BETTER?**

An alternative approach is to recognise that the effect of perturbations of the vocal tract at a particular position on the frequencies of the formants is different for the different formants because of the way they are formed in the acoustic tube. In order to understand this we must first look at how resonances are formed in the human voice apparatus and how they are affected by variation in the area of the vocal tract.

3.1. **Acoustic Theory of Speech.**

The vocal tract when saying, or singing, a vowel can be modeled as an acoustic tube, of varying area, that is closed at one end and open at the other. This gives rise to a set of distinctive standing waves, modes, resonances, or formants, as shown in figure 2 (from [11]).

![Figure 2. The first four pressure and velocity modes (formants) of a stopped pipe of uniform cross-section, from [11]. (Note: The plots show maximum and minimum amplitudes of both pressure and velocity respectively)](image)

This gives rise to a series of resonances (formants) at the following frequencies for a uniform tube $L_S$ in length.

$$F_n = \frac{nc}{4L_S}$$

Where: $c =$ The speed of sound.

$$n = 1, 2, 3, \ldots, \infty$$

$L_S =$ The tube length.

For a 17cm long uniform area vocal tract, and $c = 340\text{ms}^{-1}$ this corresponds to resonance frequencies of; $F_1 = 500\text{Hz}$, $F_2 = 1500\text{Hz}$, $F_3 = 2500\text{Hz}$ and $F_4 = 4500\text{Hz}$.

3.2. **What is the Effect of Variations in Area?**

The acoustic effect on individual pipe mode frequencies of either enlarging or constricting the size of the pipe depends directly on the mode’s distribution of standing wave pressure nodes and antinodes (or velocity antinodes and nodes respectively). The main effect of a constriction in relation to pressure antinodes (velocity nodes) is as follows from Kent and Read [12].

- A constriction near a pressure node (velocity antinode) lowers that mode’s (formant’s) frequency.
- A constriction near a pressure antinode (velocity node) raises that mode’s (formant’s) frequency.

A constriction at a pressure node (velocity antinode) has the effect of reducing the flow at the constriction because the local pressure difference across the constriction has not changed. A constriction at a pressure antinode (velocity node), on the other hand, provides a local rise in acoustic pressure, which produces a greater opposition to local airflow of the sound waves that combine to produce the standing waves. Of course the converse also applies in that the effect of locally enlarging a pipe will be exactly opposite to that of constricting it.
Figure 3. The effect of changing the area for the first three pressure and velocity modes (formants) of a stopped pipe, from [11]. (Note: The plots show the effect on both pressure and velocity respectively)

Knowing the position of the pressure and velocity nodes and antinodes for the standing wave modes in a pipe allows us to predict the effect on the mode frequencies of a local constriction or enlargement.

Figure 3 shows the potential mode frequency variation for the first three modes of a cylindrical stopped tube that would be caused by a constriction or enlargement at any point along its length. The upper part of figure 3 indicates the pressure and velocity node and antinode positions for the first three standing wave modes. The lower part of figure 3 has plus and minus signs of various sizes to indicate where that particular mode’s frequency would be raised, or lowered, by a local constriction, or enlargement, respectively, at those positions in the pipe. The size of the signs indicates the sensitivity of the frequency variation based on how close the constriction is to the mode’s pressure/velocity nodes and antinodes as shown in the upper part of figure 3.

For example, a constriction close to the closed end of a cylindrical tube will raise the frequencies of all the modes because there is a pressure antinode at a closed end, whereas an enlargement at that position would lower the frequencies of all the modes. But a constriction or enlargement at one-third of the way along the tube from the closed end would cause the frequencies of the first and third modes to be raised a little but dramatically lower the second formant.

Figure 4. Velocity nodes and antinode positions in the vocal tract for the first three modes (or formants: F1, F2, F3) of the vocal tract during a neutral non-nasalized vowel, from [11].

Figure 4 shows where the nodes and antinodes of the velocity component for the first three formants in the human voice. The first formant, and all other formants, will be significantly affected by variation of the area at the mouth where enlargement will raise their frequencies and at the larynx where enlargement will lower their frequencies. However, the middle of the tongue is a velocity node, or pressure antinode, for the second formant, so variation in area here will have a much larger effect on this formant compared to the first formant, as shown in figure 5.

Figure 5. Formant frequency modification versus position of vocal tract constriction, from [11].

Careful observation of figure 5 shows that the effect of changing the area on the different formants is position
dependent in both nature and intensity and thus can be used as a means to reduce the ambiguity of the mapping between their frequencies and the shape of the vocal tract.

3.3. Acoustic sensitivity functions

In order to make use of the varying effects changing area along the vocal tract has on formant frequencies we need to be able to relate how an area change relates to the change in frequency of a particular formant. That is how:

\[ \frac{\Delta F_n}{F_n} \Leftrightarrow \frac{\Delta a(i)}{a(i)} \] (2.)

Where: \( F_n \) = The specific formant \( \{ n = 1, 2, 3 \ldots \} \).
\( a(i) \) = The elemental tube area being varied.
\( i \) = Its position (index) along the tube.

Using (2.), we can define a sensitivity function that can be used to calculate the resultant formant frequency change, due to changes over the whole vocal tract, as follows:

\[ \frac{\Delta F_n}{F_n} = \sum_{i=1}^{N_{\text{areas}}} S_n(i) \frac{\Delta a(i)}{a(i)} \] (3.)

Where: \( S_n(i) \) = The sensitivity function.
\( N_{\text{areas}} \) = The number of elemental tube areas
\( i \) = Its position (index) along the tube.

Fant and Pauli [13] showed that the sensitivity of a particular formant frequency to a change in vocal tract cross-sectional area is related to the difference between the kinetic energy (\( KE \)) and potential energy (\( PE \)) as a function of distance from the glottis, divided by the total energy (\( TE \)) in the system. That is, the sensitivity function is given by:

\[ S_n(i) = \frac{KE_n(i) - PE_n(i)}{TE_n} \] (4.)

The total energy (\( TE \)) in the system is simply the sum of the kinetic energy (\( KE \)) and potential energy (\( PE \)) over the whole tube length.

\[ TE_n = \sum_{i=1}^{N_{\text{areas}}} (KE_n(i) + PE_n(i)) \] (5.)

The kinetic and potential energies for each formant frequency can be calculated from the pressure \( P_n(i) \) and volume velocity \( U_n(i) \) for each section of an area function as follows:

\[ KE_n(i) = \frac{1}{2} \frac{\rho(i) \lVert U_n(i) \rVert^2}{a(i)} \] (6.)

and

\[ PE_n(i) = \frac{1}{2} \frac{a(i)(i) l \rho c^2 \lVert P_n(i) \rVert^2}{i} \] (7.)

Where: \( a(i) \) = The elemental tube area.
\( l \) = The elemental tube length.
\( \rho \) = The density of air.
\( c \) = The speed of sound

As an example let us look at the sensitivity function of \( F_3 \), the third formant, from the area function for the vowel “ah”, as shown in figure 6.

![Figure 6. F3 in speech due to the area function for the vowel “ah” above.](image-url)
4. USING ACOUSTIC SENSITIVITY FUNCTIONS

The sensitivity function can be used to modify the area function, by adding, or subtracting, a scaled version of it to the area function. Direct addition of $S_1$ for any area function would raise $F_1$, whereas its opposite, $-S_1$, would lower it. $F_2$ can be similarly controlled by addition of a scaled $S_2$ replica, where $+S_2$ would increase $F_2$ and $-S_2$ would decrease it. Higher frequency formants ($F_3$, $F_4$, and $F_5$) can also be controlled via addition of their respective sensitivity functions.

Shifting multiple formants simultaneously can be carried out with addition of the sum of $\pm S_1$, $\pm S_2$, $\pm S_3$, ..., $\pm S_n$ because the sensitivity functions are superposable.

The prediction of formant frequency change based on sensitivity functions is, however, limited to small area changes (approximately 10%) as large area changes alter the sensitivity function. Thus, new sensitivity functions need to be recomputed after any small area change, and a new perturbation determined. This can be performed iteratively until arriving at an area function that produces a desired set of formant frequencies, as follows:

$$a_{k+1}(i) = a_k(i) + \sum_{n=1}^{N_{\text{formant}}} z_{n_k} S_n(i)$$

Where:
- $S_n(i)$ = The sensitivity function at time step $k$.
- $a(i)$ = The vocal tract area function at step $k$.
- $N_{\text{formant}}$ = The number of formants.
- $i$ = The position (index) along the tube.

The coefficients $z_{n_k}$ scale the sensitivity functions so that the area function perturbation moves the formant frequencies in the desired direction depending on the error. At each iteration, the $z_{n_k}$’s are determined by:

$$z_{n_k} = \alpha \left[ \frac{\mathcal{F}_n - F_{n_k}}{F_{n_k}} \right]$$

Where:
- $\mathcal{F}_n$ = The target formant frequencies.
- $F_{n_k}$ = The sensitivity function at time step $k$.
- $\alpha$ = A scale factor typically set to 10.

The iterations continue until the root of the sum of the squared differences between target formants and those of the $k$th area function, is less than some desired tolerance value.

5. PROGRAM DESCRIPTION

To make this idea work for formant frequencies extracted from the speech waveform we developed a program to do the following.

Firstly, Linear prediction is used to extract a parametric frequency model of the spectrum of the waveform.
Secondly either peak picking or root solving is used to extract formant frequencies. These are then fed to the iterative solver described above.

5.1. Formant Frequency Extraction Methods

This is relatively simple for a human to do by inspection, as it is easy for them to judge where the peaks in the spectral envelope occur, however this is more complicated to implement using software, as peak-picking algorithms would only identify the highest peaks in the signal itself, and not necessarily the envelope. We looked at two main methods one simple and the other more computationally intensive.

5.1.1. LPC with Peak Picking

There are several methods of formant detection, one of the simplest is a peak-picking algorithm applied to LPC spectrum that determines the formant. Common examples of this identify a peak if a value is larger than the previous or following values or at points where the gradient changes on either side of a value.

These algorithms are often very successful, however although they can identify clear peaks in the spectrum, they fail to recognise points that are seen as points of inflection, i.e. where there is a change in gradient but not an actual maxima.

5.1.2. LPC with Root Solving

A more reliable method of extracting the formant values is based on calculation. The smoothed spectrum used for peak picking is generated by linear prediction coefficients that represent a polynomial describing the transfer function of the signal. This polynomial can be “solved” to find its roots, i.e. the values at which the function is equal to zero in the complex plane.

The transfer function for a finite impulse response filter is a polynomial with ascending powers of $z$. It is therefore possible to find the roots of the transfer function algebraically, by transforming it into a polynomial with only positive indices and factorising to find the roots of the function. The Matlab “roots” function can be used to do this.

This method is in general more stable than the peak-picking methods, as it can identify even small peaks which might not be identified otherwise, as well as peaks that are close to other peaks, which can be seen by eye, but are not identified by a peak-picking program.

A flow chart for the program is shown in figure 8.

6. RESULTS AND TESTING

6.1. Recordings

Each of the male participants was played a C3 (approximately 130 Hz), and asked to sing each of six
vowel sounds for approximately one second each, with a break in between each one: “ah” as in Father, “eh” as in red, “ee” as in tea, “oh” as in bomb, “oo” as in boo, “i” as in fin. The one female participant sang a variety of pitches, including her lowest possible note.

The participants recorded had different levels of musical training and experience, summarised in table 1:

<table>
<thead>
<tr>
<th>Reference</th>
<th>Gender</th>
<th>Voice type</th>
<th>Vocal / musical training and ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Male</td>
<td>Medium / Baritone</td>
<td>None</td>
</tr>
<tr>
<td>M2</td>
<td>Male</td>
<td>Low / Bass</td>
<td>None</td>
</tr>
<tr>
<td>M3</td>
<td>Male</td>
<td>Medium / Tenor</td>
<td>Skilled instrumentalist but no vocal training and very little experience</td>
</tr>
<tr>
<td>M4</td>
<td>Male</td>
<td>High / Tenor</td>
<td>Some vocal training, several years in a choir</td>
</tr>
<tr>
<td>M5</td>
<td>Male</td>
<td>High / Tenor</td>
<td>Highly skilled musician, professional singer</td>
</tr>
<tr>
<td>F1</td>
<td>Female</td>
<td>High / Soprano</td>
<td>Some vocal training, several years in a choir</td>
</tr>
</tbody>
</table>

Table 1 shows details of the participants.

The participants will be referred to by their reference throughout this section.

Figure 9 shows the spectrum, spectrum, formants detected, area function, and spectrum due to the area function respectively for the singer M1 singing “ah”. The result shows good agreement for the predicted area function and similar results were obtained for the singers M2, M3 and M4. However, M5’s results were slightly different, as shown in figure 10. The formant matching to area function was good, however the initial formant frequency estimation is not very good for F5 and it seems to have underestimated this formant's frequency and this error is reflected in the area function. This may be due to the fact that there is less energy at these frequencies compared to the other males, and this may due to the fact that M5 was untrained and had a noticeably more breathy voice.
The results for the female singer at her lowest possible note G3 (approximately 196 Hz) are shown in figure 11. This vowel sound looks similar to the same sound made by the male voices, with the first two formants low and close, and the next two higher and further apart. Encouragingly, the area function also looks similar to that of all the male participants, with a narrow shape opening up to a wide mouth. In this case the formant matching program has worked reasonably well, with all the formants within about 100 Hz of their targets.

However when the female singer sang this vowel in a more normal range, at an A4, (440 Hz), A5 (880 Hz) and A6 (1760 Hz) they gave increasingly poor results. The results of her singing at a G4 (approximately 392 Hz) are shown in figure 12. Here one can see that the formant frequency estimations from the signal are very different from the lower pitch and thus the corresponding area function is very different.

As the area function was reasonable at the lower pitch this is almost certainly due to the higher pitch. At 392Hz this pitch is high enough so that the formant resonances are not sufficiently sampled for an accurate estimate from the LPC coefficients.
Figure 12 shows F1 singing “ah” at 392 Hz; spectrum, formants detected, area function, and spectrum due to the area function.

Finally figure 13 shows M3 singing “ee”. Here although the area function looks reasonable, and is consistent with the area functions of the other male singers, the actual final formant frequencies are different to those estimated from the waveform. Perhaps one of the reasons that the formant matching and area function generating programs do not work particularly well for this vowel sound is that to make an “ee” sound, the participant has to open the sides of their mouths with an action similar to smiling widely. This effectively shortens the length of the vocal tract, compared to an “ah” or “oh” sound, possibly by up to 3 or 4 centimetres, which is not taken into account by the program.

Figure 13 shows M3 singing “ee”; spectrum, formants detected, area function, and spectrum due to the area function.

For most of the vowels and male singers the estimates looked reasonable and were consistent. But the results from the one female singer were disappointing. The errors that do occur in most cases are attributable to errors in extracting the correct formant frequencies from the waveform.

7. DISCUSSION

7.1. The Effectiveness of the Method.

Unfortunately we were not able to correlate these results with some physical measures of the actual vocal tract used to make the sounds, although Story [1] has verified the accuracy for his results against MRI scans of the
vocal apparatus that made the sounds. We did investigate the output of our algorithm against a tube-based synthesis, VTDemo a "Vocal Tract Acoustics Demonstrator" [15,16] and qualitatively had good agreement. Furthermore the area functions obtained agreed with the known area functions of the vowels sung.

The main weak of the algorithm is in the accuracy of extracting the formant frequencies, which is known to be notoriously difficult to do reliably. These inaccuracies were due to both errors in the LPC coefficients as well as errors in the root extraction. Ideally the program needs to be more directly based on matching the spectra, a sort of analysis by synthesis approach, but this will almost certainly require far more computation.

7.2. The computational complexity of the algorithm.

The initial extraction of formant frequencies is based on LPC and then using either peak picking or root solving. The latter is more expensive computationally but not a significant burden for today's computing power. The area iteration is based on a self-iteration optimisation and as such is very dependent on the shape of the error surface that a particular set of formant frequencies create. Thus it is difficult to put bounds on the solution time as this varies. But we observed a running time of around 12 seconds for a frequency tolerance of 10Hz, irrespective of whether peak-picking or root solving was used. The time dropped to around 7 seconds when the frequency tolerance was increased to 50Hz. This it is almost certain that the dominant calculation is due to the optimisation routine.

For pragmatic reasons we kept the interface between the two codes as simple as possible, thus only the estimated formant frequencies were passed to the area optimisation code and the starting area function to be varied was that of the neutral vowel "a". However, as LPC can provide an estimate of the area function, it would be possible to use this estimate as the starting point for the tube solver and so improve the solution efficiency for most cases, except for those where the LPC solution is very wrong. Furthermore the optimisation was a simple iterative solver, more sophisticated methods would probably yield faster convergence and thus reduced computation times.

Also the code stressed readability and ease of modification and would benefit from significant optimisation of the code for a real time system.

8. CONCLUSIONS

This paper has presented a method for extracting vocal tract shape from the sound it produces that ameliorates the ambiguity of previous methods.

Our algorithm uses linear prediction to extract a parametric frequency model of the spectrum of the waveform and either peak picking or root solving is used to extract formant frequencies. These are then fed to an iterative solver that uses the sensitivity of these frequencies to variations in the acoustic tube area to find an optimum variation of tube area through the vocal tract that gives a minimum sensitivity solution. Both peak picking and the more computationally expensive root solver were tested for their efficacy.

Sensitivity functions work by examining the sensitivity of the vocal tract filtering effect to perturbations in its area. The different formants, or resonances, are the result of different modes of air vibration in the vocal tract, with the lowest formant frequency being a quarter wave standing wave mode, and higher formants being odd multiples of a quarter wave. Because of this the effect of perturbations on the different formants are localized to specific areas of the vocal tract and thus the ambiguity issues between vocal tract shape and the output spectrum can be resolved.

The method seems to work and produces vocal tract configurations that agree with previously measured results. However the performance of the method is limited by the quality of the formant frequency extraction and it may benefit from an alternative approach.

It not only has applications in voice training but could also be used for the control of games or music synthesis.

9. REFERENCES


