Real Time, Non-Invasive Measurements of Vocal Tract Resonances: Application to Speech Training

Annette Dowd, John Smith and Joe Wolfe [1]
School of Physics
University of New South Wales
Sydney 2052

Abstract: This study reports the determination in real time of the frequencies of the first two resonances of the human vocal tract from measurements of the acoustic impedance spectrum of the tract in parallel with the external field. The measurements were made using a broad band, frequency-independent acoustic current source, a microphone and a spectrum analyser which displays the acoustic impedance spectrum. The display provided real-time, visual feedback whereby subjects learned to imitate target vowel sounds without hearing them. Inexperienced subjects who used this feedback produced sounds that were approximately as well recognised as those produced by the same subjects imitating target vowel sounds after listening to them. The recognition rate improves with the subjects' experience in using the impedance feedback technique. This non-invasive technique could thus have possible applications in speech training and language teaching.

1. INTRODUCTION

Everyone who speaks has learned that different vowel sounds are produced by different shapes of mouth and tongue. Most adults who learn a foreign language know also that it is difficult to find the exact shape that will produce an authentic vowel sound from that language. This study reports the use of a technique[2] which measures two of the most important acoustic parameters of the vocal tract in real time. One possible application is in speech training and language teaching.

Phoneticians arrange vowels in a space whose two dimensions are the mouth opening (negative y axis) and the horizontal position of the tongue (x axis) [1]. Acousticians arrange vowel sounds in a two dimensional space whose dimensions are the frequencies of the first two formants [2]. With suitable choice of directions of axes, the relative positions of the vowels in these two representations are the same [3]. This similarity is usually explained thus: the formants in the sound are due to the first two resonances of the vocal tract, and the frequencies of these resonances are largely determined by mouth opening and tongue position [4]. The first two formants are traditionally called called F1 and F2. We shall call the first two resonances R1 and R2 in order to distinguish between formants (features of the sound) and resonances (features of the tract that produces it).

Direct measurement of the resonances of the vocal tract has both intrinsic and practical interest. The frequencies of resonance of the tract are functions only of properties of the tract (its geometry and the mechanical properties of the air it contains and its walls) and they can be measured and defined relatively precisely. The frequencies of formants in voiced speech are more difficult to measure precisely because the output sound depends on the waveform input at the glottis as well as on the transmission and radiation properties of the tract. Sundberg has measured the resonances by mechanical excitation at the throat and measurement of the radiated sound [5], and Badin and colleagues have used this technique to measure the transfer function [6,7].

The principal aim of this paper is to investigate whether real-time, non-invasive measurements of the first two resonances can be used to provide a visual feedback in speech training. Such feedback could be useful to those who have inadequate auditory feedback. One case is those with severe hearing impairment. Another is that of adults learning foreign languages: such students often find it difficult to distinguish between the the target sound (e.g. the sound of a vowel as produced by a native speaker) and their attempt to imitate that sound. This problem is attributed in part to categorisation: a listener who has learned to divide speech sounds into the finite categories of his/her own language has a tendency to divide sounds in a new language according to those same divisions [8].

Why is it necessary to measure R1 and R2 to provide visual feedback? If the object is to imitate a sound, why not just imitate the waveform? While it is true that the same waveform implies the same vowel, it is not true that two examples of the same vowel must have similar waveforms. This is especially the case if the sounds have different pitch. Figure 1 shows five oscillograms. Four are examples of the same vowel pronounced with different pitch and loudness (a,b,c,d). The fifth (e) is a different vowel. Of these five, the two which appear most similar in shape are (a) and (e). Further, it is not a simple task to distinguish vowels from inspection of the spectrum alone. The problem in this case is that the spectrum of a voiced vowel with fundamental frequency f_0 has frequency components f_0, $2f_0$, $3f_0$ etc. In

[1] To whom correspondence should be addressed

[2] This technique is the subject of a patent application PCT/AU95/00729

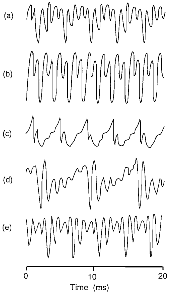

Figure 1. Oscillograms of five vowels with arbitrary amplitude. 1a, 1b and 1c were spoken by the first author (a soprano). 1a is that of the vowel /ɑ/ (as in "hard") at normal conversational pitch and volume. 1b is that of /ɑ/ at higher pitch and normal volume. 1c is that of /ɑ/ at normal pitch and lower volume. 1d is that of /ɑ/ spoken by the third author (a bass). 1e is that of the vowel /æ/ (as in "had") spoken by the first author at normal conversational pitch and volume.

the case where $f_0 << R1$ (a bass voice in its low register), the sampling of the spectral envelope is usually adequate to estimate R1 and R2 reliably. For high pitched voices, including those of children, f_0 is often of the same order as R1, or even higher, so the sampling of the spectral envelope is not adequate for the identification of R1, and sometimes even R2 cannot be estimated accurately[3].

This study reports the determination of the frequencies of the first two resonances of the human vocal tract from measurements of the acoustic impedance of human vocal tracts in parallel with the external field. The measurements were made from a position outside but near the lower lip, using a broad band acoustic current source, a microphone and a spectrum analyser which displayed the results in real time (< 100 ms). This display then provided visual feedback. The system was then used as a speech trainer, and its performance was compared with those of two other forms of feedback (Figure 2). In one series of trials, a photograph of the lower face of the speaker pronouncing the target vowel was showed to subjects who were then asked to imitate the mouth geometry of the speaker and to phonate (this protocol conveys the information about the vowel that is conveyed in speech reading, also called lip reading). In another series, the subjects heard a recording of the speaker pronouncing the target vowel and were asked to produce the same vowel. This protocol is a model of the feedback usually available to those learning to speak.

2. MATERIALS AND METHODS

2.1 Acoustic impedance

Acoustic impedances were measured using a technique described in detail previously [10,11]. Briefly: a computer produces a periodic waveform comprising the sum of many sine waves. This is output via a digital-analog converter to an amplifier and then to a loudspeaker. On one side of the speaker is a sealed enclosure; on the other is an impedance matching horn which connects it acoustically to an annular acoustic resistor. This resistor has an acoustic output impedance of 33 MRayl which is larger than the loads measured (typically 1 MRayl), so the source approximates an ideal current source. The sound source (the output of the acoustic resistor) is located next to a microphone in a measurement head. The microphone measures the acoustic pressure produced and its signal is input to a spectrum analyser. The digitised waveform output by the computer is calculated during a calibration procedure using a resistive load. The coefficients of the sine terms in the synthesized waveform are calculated so as to compensate for the frequency dependence of the amplifier, loudspeaker, impedance matching horn, output impedance and microphone, so the output acoustic current has the same amplitude for all frequency components. The pressure spectrum measured by the microphone is therefore proportional to the acoustic impedance at the measurement head.

A finite signal to noise ratio and a finite measurement time together impose a maximum information content in any measurement. A compromise must therefore be made for the acceptable values of frequency range, frequency resolution, amplitude resolution and measurement time. The signal to noise ratio in these experiments was limited by three constraints. First, the digital analog card used was only 12 bit. Second, the measurements were made in a room with a background noise of typically 45 - 55 dBA. Third, the external field is a low value acoustic impedance in parallel with the vocal tract. We decided not to improve these conditions for this study because a practical speech trainer ought be simple, non-invasive and able to work in a classroom environment.

[3] This raises the obvious objection that, since human ears and brains can usually identify vowels from the sound, even when spoken by high-pitched voices, it should be possible for a measurement system to do so as well. There are two obvious responses. First, human ears and brains do sometimes have problems in identifying vowels with a constant high pitch. Second, human brains have the added information of linguistic and syntactic context, and the relatively high redundancy of human speech [9]. A listener who is familiar with English will know which of the the syllables "hud" and "had" is intended by an English speaker because the first is not an English word. A listener will know which of "he heard me" and "he head me" is intended because the latter is not an English sentence.

The frequency range was chosen as 200 - 2500 Hz, with a resolution of 25 Hz. Thus the output current was a periodic waveform produced by summing 93 sine waves with frequencies 200, 225, 250... 2500 Hz. The microphone power spectrum was displayed using a Spectral Innovations MacDSP card in a Macintosh II computer implementing a Fourier transform with a sampling rate of 15.62 kHz and a Hamming window. The microphone and the output of the acoustic resistor were mounted 10 mm apart in a block of nylon used as a measurement head. This head was placed against the subject's lower lip for vocal tract measurements. For calibration, the measurement head was used to seal one end of a 35 m tube. During a measurement lasting less than 200 ms, the reflection does not return from the distant end of this tube, so it is effectively infinite and therefore a resistive load [13].

2.2 Preliminary subject training

For most people, the relaxed position of the vocal tract at rest (i.e. when not phonating) has the velum (soft palate) lowered towards the dorsal aspect of the tongue. It was therefore necessary to show subjects how to hold the vocal tract in the phonating position while not phonating. Two different methods of feedback were used to teach this technique. Using a mirror, subjects found that they could see the palate in the relaxed position, and that when they phonated (saying "ah"), it lifted and they could see the backs of their throats. They were then asked to practise lifting the palate without phonating. The other method used the measurement of acoustic impedance: with the palate raised, two resonances were usually observed in the range of the impedance spectrum; with it lowered, only one resonance was observed. Most volunteers learned this skill in about 10 minutes, although some did not learn in 30 minutes and were not included in the study. More volunteers over the age of 25 (75%) than under this age (15%) were unable to learn this skill in 30 minutes.

2.3 "Target" vowel sounds and "target" impedance spectra

Two speakers (one male and one female) were chosen to produce vowel sounds and corresponding impedance spectra which were to be imitated by the test subjects. The male speaker was 22 years of age and had lived most of his life in the North of Sydney and has a mild Australian accent. The female speaker was 21 years of age and had lived the last 10 years in the North of Sydney. Her accent is predominantly Australian, but with a slight trace that suggests she is not a native English speaker. Both have university educations.

Nine vowels were chosen for this study: /e/ as in "head"; /ɜ/ as in "heard"; /ɑ/ as in "hard"; /æ / as in "had"; /ʌ/ as in "hut"; /ɒ/ as in "hot"; /ɔ/ as in "hoard"; /ʊ / as in" hood"; /u/ as in "who'd". For each vowel, the speakers pronounced a sustained vowel which was digitised and recorded by a Macintosh computer. Immediately after they ceased phonating, an impedance spectrum was measured. This was repeated several times and the means and variances in the frequencies of the first two resonances were noted. The mean values were used as the target values, and the target sound was the vowel sound recorded immediately before the impedance spectrum whose resonances most closely approached the mean values. A photograph was taken of the lower half of the face of each speaker pronouncing each (sustained) target vowel.

2.4 Vowel imitation tests

Subjects were asked to imitate the nine target vowels using three different feedback methods.

i) **Photograph only.** This was designed to convey, under controlled conditions, the information that is available when speech reading ("lip reading"). During these tests, the subjects wore headphones playing white noise to mask other sounds and so to minimise the subjects' use of the sound of their own voice as a clue to modify the sound they were producing. The subjects were shown the photograph of the lower half of the speaker (of the same sex) pronouncing the target vowel. They were instructed to imitate the mouth position of the speaker and to phonate. This sound was recorded on a cassette recorder. An impedance spectrum was measured immediately following the phonation. This procedure was repeated three or four times for each vowel. (Figure 2a.)

ii) **Impedance spectrum plus photograph.** During these tests, the subjects wore headphones playing white noise to mask other sounds and so to minimise their use of the sound of their own voice as a clue to modify the sound they were producing (Figure 2b). The measurement head was placed at the subject's lower lip, and the impedance spectrum was continuously displayed on a monitor in the subject's view. The subjects were permitted to practise varying the frequencies of the two resonances in acoustic impedance, Z(f), and were told that opening the mouth raised R1 and that bringing the tongue forward raised R2. For each target vowel, a transparent sheet was placed over the monitor screen, positioned so that two vertical lines drawn on the transparency indicated the frequencies of the first two resonances of the target vowel, as articulated by the target speaker of the same sex. The photograph of the lower half of the speaker's face was also displayed. Subjects were instructed to use the photograph as a starting position, and then to move tongue and mouth such that the local maxima in Z(f) coincided with the target values. When the subject was satisfied that s/he could not easily improve the degree of match, s/he phonated and the sound was recorded on cassette tape. Z(f) was stored and the resonant frequencies noted. Each subject had three or four attempts at each vowel.

iii) **Auditory.** For these tests, the subject wore headphones through which were played the recordings of the target vowels pronounced by the speaker of the same sex as the subject (Figure 2c). The subject could listen to each vowel as many times as s/he liked, and could hear his/her own voice and thus adjust the sound to match that of the target. When s/he was satisfied with the match, the phonated sound was recorded and, immediately afterwards, an impedance spectrum was measured.

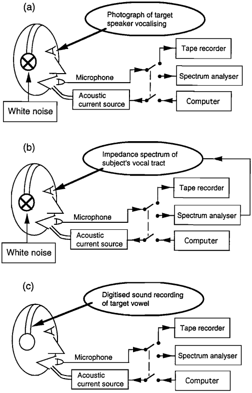

Figure 2. A schematic representation of the three different protocols used for the imitation of vowel sounds. In the first (2a) the subject sees only a photograph of the lower face of the target speaker. In the second (2b) s/he sees both the photograph and the impedance spectra of his/her own vocal tract, upon which is superimposed the frequencies of the resonances of the vocal tract of the speaker when producing the target sound. In the third (2c) the subject sees nothing, but hears a recording of the speaker pronouncing the target vowel.

2.5 Subjects

Nine subjects (six males and three females) each recorded at least once the complete nine vowel set using the three different feedback methods. 350 spectra and 470 vowel sounds were recorded. 5% of these were discarded because of either poor quality recording (extremes of sound level or wind noise), or the inability of the subject to maintain constant position with his/her tongue and lips.

Six speakers were born outside of Australia. Of these, four learnt a language other than English as their first language. This was not a deliberate attempt to include more variables, but rather the result of finding volunteers in a multi-cultural society. Two of the subjects did the complete set of imitations of all vowels twice. One of these subjects was a female, 13.5 years old, whose native language is Russian and who has lived in Sydney for 2.5 years and has an accent that is predominantly Australian, but with a trace indication that English is not her native language. The other was male, a native English speaker, had lived in Australia for 40 years and has a mild Australian accent. The data and recordings of these subjects were used for further analysis to investigate improvements in imitation between first and second trial.

2.6 Listening panel

The sounds recorded by the reference speakers, and the imitations made by the two subjects who completed the series twice (135 sounds in total) were transferred in random order onto a cassette tape. This tape was played independently to each member of a listening panel of six native English speakers (aged from 22 to 46 years). The listening panel were given a list of nine vowels described as the vowel sounds in the words "head", "heard", "hard", "had", "hut", "hot", "hoard", "hood" and "who'd". The sounds were each played four times, and the listeners could replay any sound if desired. They were asked to decide which of the nine listed vowels was most closely approached by the sound they had just heard. If undecided, they were instructed to leave that entry blank on the response form.

3. RESULTS AND DISCUSSION

Figure 3a shows the pressure spectrum recorded when the measurement head was connected to the "infinite" tube after calibration. The output signal of the digital-analogue converter has been adjusted to reduce the frequency dependence of the pressure spectrum to less than 0.1 dB r.m.s. into such a resistive load [10]. The acoustic current is therefore also independent of frequency. Further, because the output resistance of the source is high, this current is independent of external loads with low impedance. The measured pressure spectrum is therefore approximately proportional to the acoustic impedance of a load connected to the measurement head. The 25 Hz "ripple" in the nominally flat spectrum is due to the sampling frequency. This low resolution in frequency was a necessary compromise to allow measurements in the noisy, low-impedance laboratory field.

Figure 3b shows the measured pressure spectrum with the measurement head positioned next to the lower lip of a subject with his mouth closed. In a free field, the acoustic impedance Z(r) at a point r from an isotropic source is $\rho cjkr/(1 + jkr)$ where ρ is the density of the medium, c the speed of sound, k the wave number and $j = \sqrt{(-1)}$. The measurement head is much smaller than 1/k for the wavelengths considered here, so $r << 1/k$. For most of the frequencies used here, the dimensions of the subject's head are comparable with or larger than the wavelength, so his face acts as a baffle. Thus the sound is radiated into approximately 2π steradians, rather than 4π, which approximately doubles the impedance. For a small measurement head and an infinite baffle $Z(r) \approx 2\rho cjkr$. This condition is roughly approximated by the experimental arrangement, which explains the rise of about 3 dB/octave in the measured impedance spectrum.

Figure 3c shows the spectrum measured at the lower lip of the male target speaker with his tract prepared to pronounce /ɒ/ (as in "hot"). This is a measurement of the impedance of

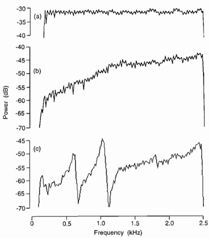

Figure 3. The measured pressure spectrum using the measurement head with three different loads. 3a shows the measured spectrum when the head is connected to an "infinite" tube for calibration. 3b shows the spectrum measured at the lower lip of a subject in the laboratory field, which his mouth closed. 3c shows the spectrum measured at the lower lip of the male target speaker with his vocal tract prepared to pronounce the vowel /ɒ/ (as in "hot").

his vocal tract in parallel with that of the field of the laboratory. The latter is a low impedance shunt which dominates the measurement except at resonances of the vocal tract. Resonances are seen at 0.6 and 1.0 kHz. Note that the impedance rises and falls below and above the resonance frequency. The lab field impedance, including the baffle, is inductive. The impedance of the tract is capacitive below resonance and inductive above resonance, so the parallel impedance is respectively higher and lower than that of the lab field. There is some noise present at around 200-300 Hz. This noise was one of the reasons that led us to exclude the vowels /i/ and /I/ from this study. In Australian English, these vowels have R1 in the range 200-350 Hz. Further, the low and wide mouth opening associated with them means that relatively little acoustic energy from the measurement head enters the mouth. This low signal and the high noise in the relevant frequency range meant that we could not reliably determine R1 for these two vowels with the current version of the apparatus.

Figure 4 shows plots of the resonances (R1,R2), as determined from the impedance measurements, for the nine vowels studied for the female subjects only. The (R1,R2) of the target vowels are indicated by the heads of the arrows. Figure 5 shows the analogous data for male subjects only. Subjects attempted to imitate the vowels of the target speaker of the same sex. Flanagan [12] reports that the adult female vocal tract is on average 0.87 times the length of the male vocal tract, and this difference affects the range of formant frequencies. The values for (R1,R2) in Figure 5 are similar to, but not identical to, the mean values of (F1,F2) reported by Bernard [14] for 100 Australian speakers of English using a spectrograph. In all cases, the shaded ellipses represent the attempts by the subjects to imitate the target vowels. The centre of each ellipse is located at the mean values of R1 and R2 for the imitation of that vowel. The horizontal and vertical semi-axes of each ellipse are the standard deviations in the R1 and R2 of the imitations.

Figures 4a and 5a show the (R1,R2) of the imitations made by the subjects using only the photograph of the target speaker. As might be expected of a method that gives little information about the position of the tongue, the data are rather scattered. There is less scatter in R1, which is determined predominantly by mouth opening and less by tongue position, and more scatter in R2, which is determined largely by the front-back position of the tongue, a parameter which cannot easily be determined from the photograph. All of the subjects spoke English fluently, and all knew that they were imitating phonemes in Australian English. This

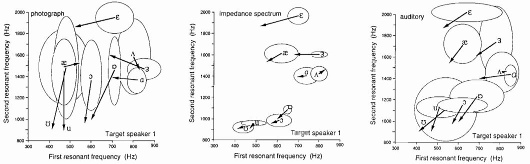

Figure 4. The vocal tract resonant frequencies (R1,R2) for the target vowels spoken by target speaker 1 (indicated by the heads of the arrows) and the mean values of (R1,R2) for the imitations (tails of the arrows) by the subjects using three different feedback protocols. Figure a shows photograph only, b shows photograph plus impedance spectrum, c shows auditory feedback. The horizontal semi-axis of each shaded ellipse is equal to the standard deviation of R1 in the imitations, and the vertical semi-axis is that in R2. Speaker 1 was female and the data are those of the female subjects imitating her target vowels.

/ɛ/ as in "head"; /ɜ/ - "heard"; /ɑ/ - "hard"; /æ/ - "had"; /ʌ/ - "hut"; /ɒ/ - "hot"; /ɔ/ - "hoard"; /ʊ/ - "hood"; /u/ - "who'd".

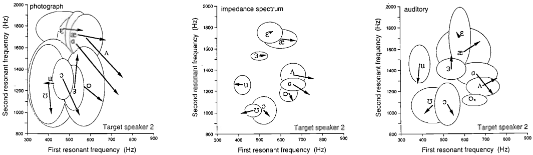

Figure 5. As for Figure 4, but for target speaker 2 (a male). The data are for the male subjects. /ɛ/ as in "head"; /ɜ/ - "heard"; /ɑ/ - "hard"; /æ / - "had"; /ʌ/ - "hut"; /ɒ/ - "hot"; /ɔ/ - "hoard"; /ʊ / - "hood"; /u/ - "who'd".

information may have resulted in smaller scatter than would have been the case if the photographs were the only information. For instance, the vowel /ɛ/ has an unambiguous mouth shape in Australian English and the subjects may have known that only one tongue position is associated with that mouth shape.

Figures 4b and 5b show the (R1,R2) of the imitations made using the impedance spectrum together with the photograph of the target speaker. In this protocol, and in that using the photograph only, the subjects wore headphones emitting white noise which masked any noise they made themselves, and so they were unable to use their own voice as feedback. The scatter in both R1 and R2 is much less than in Figures 4a and 5a. In contrast with Figures 4a and 5a, the scatter in R2 was less than that in R1. Subjects commented that it was easier to match R2 than R1. This may be because R2 varies over a greater range, and is associated with a larger change in Z(f). Thus the variation in R2 was more noticeable in the spectrum and this may have encouraged the subjects to concentrate more on matching it than on matching R1. This feedback protocol shows the least scatter of the three used. Note however that the scatter is always rather greater than 25 Hz, the resolution of the impedance spectrum displayed. We suspect that this was due to the finite skill and patience of the subjects in matching the frequencies (discussed further below).

Figures 4c and 5c show the (R1,R2) of the imitations made using auditory imitation: the subjects listened to a recording of the speaker pronouncing the target vowel and then made as many attempts as they wished to produce the same sound. In this protocol there was no masking of the sound of the subject's own voice.

In order to summarise these data, we define a parameter to indicate the difference between the imitation and the target resonances. Displacement in the (R1,R2) plane is an obvious measure, but this gives extra weighting to the imitation of the second resonance because its range is larger than that of the first. For this reason, we weight the R1 and R2 components of the displacement with the reciprocals of the standard deviations σ_1 and σ_2 of R1 and R2 in all nine vowels used. The scaled displacement d in the (R1,R2) plane is therefore defined as:

$$d = \sqrt{\left(\frac{R1 - R1_t}{\sigma_1}\right)^2 + \left(\frac{R2 - R2_t}{\sigma_2}\right)^2} \qquad (1)$$

where the subscript t indicates the target value. Using this definition, a perfect match has d = 0, and a poor match may have d as high as 3. In the data of Peterson and Barney [2], the range of d for a typical vowel in the population studied is about 1, so a d of this order might be considered an acceptable match.

Table 1 shows the scaled displacement (d) for all vowels using all methods for both sexes. It is not surprising that imitation using the impedance spectrum gives the smallest displacements in the (R1,R2) plane: after all it is precisely the displacement of R1 and R2 that the subjects are seeking to minimise, and the scatter here is presumably just a measure of the limits of their skill and patience. It is perhaps more surprising that the scatter is so large for auditory imitation, given that all of the subjects have considerable experience in this feedback and were all presumably highly skilled at this when they learned to speak. This may, in part, be the result of the process of categorisation in the mature auditory system: an adult who has learned one language tends to divide the (R1,R2) plane into the vowels of that language, and fails to discriminate small differences. For example, one subject produced exactly the same (R1,R2) when imitating the target sounds /u / and /æ /. The /u/ was a good match, but the /æ / a poor one. This can be explained if he perceptually categorised the vowel as /u/ in both cases, and then produced his version of that vowel, rather than a good imitation of the target. (There are several other such examples in the data.) It is possible that much of the scatter is due to categorisation: once a subject categorises the vowel s/he hears, s/he then pronounces his/her usual version of it rather than a faithful imitation of the target vowel. Although several of the subjects in this study spoke two languages with reasonable fluency, none regarded themselves as particularly good at languages, and none had specialised voice training.

Two of the subjects repeated the imitation trials. Table 2 shows their average d for the first and second session using each method of feedback. With the photograph only and with auditory feedback, there was no significant improvement

TABLE 1. Scaled displacement (d) for all imitations

Vowel	/ɛ/ *"head"*	/ɜ/ *"heard"*	/ɑ/ *"hard"*	/æ/ *"had"*	/ʌ/ *"hut"*	/ɒ/ *"hot"*	/ɔ/ *"hoard"*	/ʊ/ *"hood"*	/u/ *"who'd"*	**Average:** **all vowels**
(photograph,only)	1.24	1.53	1.91	1.72	1.51	0.92	1.40	1.51	1.59	1.48 ± 0.28
(impedance,feedback)	0.60	0.83	0.66	0.92	0.71	0.43	0.53	0.46	0.54	0.63 ± 0.17
(auditory,feedback)	1.30	1.17	1.23	1.42	0.92	0.75	0.98	1.09	1.07	1.10 ± 0.19

between the first and second attempts. With the impedance feedback, however, the improvement was significant at 95%. This could be because impedance provides the most novel feedback - we presume that all of the subjects had learned to speak using sounds and mouth shapes of speakers as targets. Despite this novelty, the goodness of the match in the second attempt suggests that this technique can be easily learned, which is an important feature for the use of the technique as a speech trainer.

TABLE 2. Scaled displacement (d) for first and second trials

	first session	second session
(photograph,only)	1.60 ± 0.90	1.36 ± 0.76
(impedance, feedback)	0.73 ± 0.39	0.47 ± 0.41
(auditory, feedback)	0.86 ± 0.50	1.03 ± 0.57

The data in Tables 1 and 2 suggest that the impedance feedback technique makes a very good speech trainer for vowels - perhaps even better in this respect than the traditional auditory feedback. It could be argued however that this comparison is unfair, in that the value d quantifies the very parameters that are minimised in the impedance feedback technique. What is important in vowel pronunciation is not the values of R1 and R2 *per se*, but rather how the vowel sound is interpreted by a listener. For this reason we conducted listening tests on the recordings of the phonations made by the subjects using the three techniques. Members of the listening panel listened to a cassette tape containing the target vowels and all of the imitations of them in randomised order. The panel members were given a list of the nine target vowels and were asked to note, for each sound they heard, which of the vowels it most resembled, or to leave the example blank if undecided. The raw data were then assembled in confusion matrices: tables with the target vowels in columns and the perceived vowels in rows[4]. The data were then summarised as a success rate (Table 3). The success rate for the target speakers is significantly different from all others at 95% confidence level, as is that from the photograph only. The success rates of the sounds produced using impedance feedback and auditory feedback are not significantly different.

TABLE 3. Identification from listening tests

Feedback protocol	(Percentage correctly, identified by listeners)
Photograph only	29 ± 6%
(Impedance spectrum)	39 ± 9%
Auditory feedback	40 ± 9%
Target speakers	61 ± 10%

A random choice with nine vowels would give a success rate of 11%. The imitations using a photograph only showed the lowest success rate: 29%. Most of the vowels were perceived as either /ɛ/ or /ɜ/ by the panel. This is difficult to explain, because /ɛ/ and /ɜ/ have similar R1 and differ in R2 in Australian English. They are however close to the middle of the (R1,R2) plane and are therefore "neutral" vowels. It is possible that the subjects, not having information about tongue position, automatically put it in a middle position vertically and longitudinally, leading to neutral vowels.

The vowels spoken by the target speakers gave the highest success rate: 61%. This seems surprisingly low at first, given that we are able to identify correctly a greater percentage of words. Part of the difference comes from the context. For example, in Australian English /ɑ/ and /ɒ/, pronounced in isolation, might be difficult to identify (see Figure 3 or [14]), but "hard" and "hod" are more readily distinguished if one knows that the sound is an English word. In the classic study by Peterson and Barney [2] on American vowels, listeners identified vowels with a success rate of 94% when they were presented in the context of syllables beginning with "h" and ending with "d". In this study, the vowels were presented without beginning and ending consonants, and this unfamiliar context may have deprived listeners of clues about the vowels that come from familiar transients. Vowel duration is also important [14]: /ɑ/ and /ʌ/ have very similar formant frequencies, but the words "heart" and "hut" are readily distinguished by vowel duration. The loss of information about vowel duration was probably important in this study, because all vowels were pronounced in sustained form. For instance, the listeners identified /ʌ/ less frequently than other vowels, presumably because they did not recognise it when presented as a sustained vowel.

Identification of the sounds produced by auditory imitation gave a success rate of 40%. This low value also seems surprising at first. In this case, all the problems mentioned in regard to the target vowels apply, and there is the further problem of categorisation. This protocol involves two perceptions of the vowel: once by the subject and once by the listening panel. At each stage the the opportunity for a categorisation error arises. This is supported by the observation that the success rate for auditory imitation is approximately the square of the success rate for the target vowels themselves.

Identification of the sounds produced using the impedance spectrum as feedback had a success rate of 39%. It is a little surprising that this novel and unfamiliar feedback technique

[4] The raw data for this study are presented by Dowd [15] in an undergraduate thesis. This thesis contains other data, such as measurements of an American speaker, measurements of the formants using other techniques and use of the impedance technique to measure the impedance of model systems for which the formants may be calculated analytically. Copies of this thesis can be supplied upon request.

was approximately as successful as auditory feedback. Further, those subjects who returned for a second session showed improvement both in the speed of production of a sound imitation and in the recognizibility of the sounds produced. From the first to second session the rate of recognition increased from 36 ± 10% (18) to 41 ± 9% (18) (significant at 90%). In contrast, the second session using the photograph only or auditory feedback showed no significant improvement in scaled displacement or recognition rate. These results are consistent with those of the scaled displacement.

4. CONCLUSIONS

Our non-invasive technique demonstrates the possibility of measuring the impedance spectrum of the vocal tract in parallel with the external field in real time. With the exception of /i/ and /I/, both of the first two resonant frequencies for vowels may be seen sufficiently clearly to allow the impedance spectrum to provide feedback in a speech trainer with minimal training of subjects. Inexperienced subjects who use this feedback to imitate target vowels produce sounds that are approximately as well recognised as those produced by the same subjects listening to and imitating vowel sounds. The recognition rate improves with the subjects' experience in using the impedance feedback technique.

ACKNOWLEDGEMENTS

We thank our volunteer subjects for their time and patience.

REFERENCES

1. D. Jones, *An Outline of English Phonetics.* (Ninth ed.) Heffer & Sons, Cambridge, (1960)
2. G.E. Peterson and H.L. Barney "Control methods used in a study of vowels" *J. Acoust. Soc. Am.* **24**, 175-184 (1952)
3. M. Joos, "Acoustic phonetics" *Language* **24**, 1-136 (1948)
4. J. Clark and C. Yallop. *An Introduction to Phonetics and Phonology*, Basil Blackwell Ltd, Oxford (1990)
5. J. Sundberg, *The Science of the Singing Voice*, Northern Illinois Univ. Press., De Kalb, Ill. (1987)
6. Y. Pham Thi Ngoc and P. Badin. "Vocal tract acoustic transfer function measurements: further developments and applications" *J. de Physique* **IV**, C5, 549-552 (1994)
7. Y. Pham Thi Ngoc, *Caractérisation acoustique du conduit vocal: fonctions de transfert acoustiques et sources de bruit*, Doctoral thesis, Institut National Polytechnique de Grenoble (1992)
8. A. Landercy and R. Renard, *Éléments de Phonétique*, Didier, Bruxelles (1977)
9. N.H. Fletcher, *Acoustic Systems in Biology*, Oxford, New York (1992)
10. J. Wolfe, J. Smith, G. Brielbeck and F. Stocker. "Real time measurement of acoustic transfer functions and acoustic impedance spectra" *Australian Acoustical Society Conference*, Canberra, pp. 66-72 (1994)
11. J. Wolfe, J. Smith, G. Brielbeck, and F. Stocker "A system for real time measurement of acoustic transfer functions" *Acoustics Australia*, 23, 19-20 (1995)
12. J.L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag, Berlin (1972)
13. N.H. Fletcher and T.D. Rossing, *The Physics of Musical Instruments*. Springer-Verlag, New York (1991)
14. J.R.L. Bernard, "Length and identification of Australian vowels", *J. Australasian Universities Modern Language Assoc.* **27**, 37-58 (1967)
15. A. Dowd, *Real Time Non-Invasive Measurements of Vocal Tract Impedance Spectra and Applications to Speech Training*, Undergraduate thesis, School of Physics, University of New South Wales, Sydney (1995) Copies available upon request.