Meas. Sci. Technol. **8** (1997) 1112–1121. Printed in the UK PII: S0957-0233(97)81349-5

A novel instrument to measure acoustic resonances of the vocal tract during phonation

J Epps, J R Smith and J Wolfe

School of Physics, University of New South Wales, Sydney, New South Wales 2052, Australia

Received 28 January 1997, in final form 27 May 1997, accepted for publication 1 July 1997

Abstract. Acoustic resonances of the vocal tract give rise to formants (broad bands of acoustic power) in the speech signal when the vocal tract is excited by a periodic signal from the vocal folds. This paper reports a novel instrument which uses a real-time, non-invasive technique to measure these resonances accurately during phonation. A broadband acoustic current source is located just outside the mouth of the subject and the resulting acoustic pressure is measured near the lips. The contribution of the speech signal to the pressure spectrum is then digitally suppressed and the resonances are calculated from the input impedance of the vocal tract as a function of the frequency. The external excitation signal has a much smaller harmonic spacing than does the periodic signal from the vocal folds and consequently the resonances are determined much more accurately due to the closer sampling. This is particularly important for higher pitched voices and we demonstrate that this technique can be markedly superior to the curve-fitting technique of linear prediction. The superior frequency resolution of this instrument which results from external vocal tract excitation can provide the precise, stable, effective, articulatory feedback considered essential for some language-learning and speech-therapy applications.

1. Introduction

The sustained sounds of voiced speech, including vowel sounds, are nearly periodic signals and their spectra comprise the fundamental frequency of vibration of the vocal folds ('vocal cords') and integral multiples of that frequency. In most European languages, information about the individual speech sounds (phonemes) is carried by the spectral envelope and is largely independent of the pitch (that is, the fundamental frequency). Most of the information about vowels is contained in the local maxima in the spectral envelope (formants) produced by resonances of the vocal tract. (The term 'formant' is sometimes used to refer both to the peak in the spectral envelope and to the resonance responsible for that peak. In this paper we maintain the distinction and only refer to the peaks in the spectral envelope as formants.) The frequencies of the resonances are functions of the shape of the vocal tract, especially of the mouth opening and the position of the tongue (reviewed by Clark and Yallop (1990)).

Precise measurements of the frequencies of resonance are therefore of interest in acoustical phonetics. They can also supply feedback about mouth shapes and tongue positions for applications in speech therapy and language learning. The resonance frequencies are usually estimated from the speech signal itself, but the precision of such estimates cannot be very much better than the harmonic spacing, namely the pitch frequency of the voice. This precision can be inadequate, particularly in cases in which the pitch frequency is comparable to or greater than the resonance frequency of interest. The resonance frequencies of high-pitched voices (such as those of children and some women) are thus especially difficult to determine.

This paper introduces a novel instrument that can precisely measure the resonance frequencies of acoustic systems in real-time in the presence of an interfering harmonic signal and describes how it can be used to study the resonance frequencies of the vocal tract during normal phonation. It employs a non-invasive technique (real-time acoustic vocal tract excitation or RAVE) that involves exciting the vocal tract just outside the lips with a broadband acoustic current source and suppressing the speech signal component of the measured pressure spectrum. The resulting magnitude response is then used to calculate the resonance frequencies. In one particular application presented herein, the two lowest resonances are displayed as the coordinates of a moving cursor in two dimensions.

0957-0233/97/101112+10$19.50

2. Techniques for measuring vocal tract resonances

2.1. Estimation from formants of normal speech

Speech production is often modelled as a source and filter (Fant 1960). The source is the periodic, harmonic-rich pressure wave produced when air flow from the lungs is modulated by the vibrating vocal folds. The vocal tract (which extends from the vocal folds to the lips) is considered as a time-varying filter whose acoustic gain is frequency dependent due to the resonances produced by its physical geometry. When the fundamental or pitch frequency is small compared with the spacing of these resonances, the output speech signal carries a broad band of acoustic power (formant) for each of these resonances (see figure 1). Up to five distinct formants can sometimes be seen in the speech spectrum, and each is associated with a vocal tract resonance. The vocal tract resonances in turn are associated with the configuration of the vocal tract. Vowel sounds are associated with the three resonances of lowest frequency, but may usually be identified by the two resonances with lowest frequency (Landercy and Renard 1977, Clark and Yallop 1990). These two resonance frequencies are largely determined by the mouth opening and the position of the tongue.

Vocal tract resonances give rise to the formants present in speech signals measured outside the lips; thus the formant frequencies (when they can be determined) are assumed to approximate the resonance frequencies. Since the introduction of spectrograms, phoneticists, speech therapists and others have inspected spectrograms of the speech signal to estimate the formant frequencies. The involvement of a human to interpret the spectrum introduces subjectivity and other potential artefacts into the measurement process and also limits the speed of the process. Furthermore, the harmonic spacing of the signal which excites the tract imposes a fundamental limit on the precision of this method. For an adult male speaker with a pitch frequency below 100 Hz (a rather low-pitched voice) an error of order ± 50 Hz may be considered acceptable for some applications. For a child or soprano speaking with a fundamental frequency of say 300 Hz or higher (Sundberg 1987), the inaccuracy is more serious, see figure 1.

This observation raises the following question: how important is it to obtain the formants or resonances precisely? After all, human listeners can usually identify speech sounds from vocal tracts which are excited only by the speech signal. If humans can do it, why should a feedback device need more precision? Human recognition of speech, however, is performed by neural networks with many years of training and, furthermore, uses the extra clues provided by syntactic and linguistic context. Highly accurate word recognition is attainable because of the high level of internal redundancy in language (Fletcher 1992). To give an example, it is often possible to identify words accurately _v_n wh_n v_w_ls _r_ c_mpl_t_l_ m_ss_ng. A non-human feedback device will lack these contextual clues and consequently will require more information about the vocal tract configuration, namely a more precise measurement of the resonance frequencies.

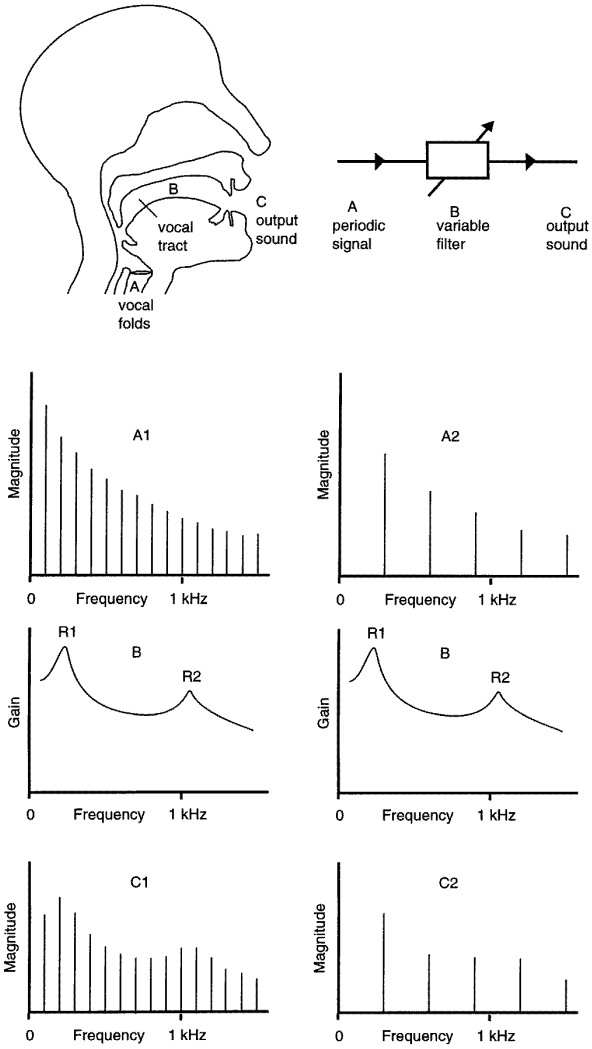

Figure 1. The process of speech production, showing idealized spectra for a bass male voice (1) and a soprano female voice (2). Voiced speech is commonly analysed as a periodic signal (A) from the vibrating vocal folds, which is input to the vocal tract (B) which acts as a variable filter to produce the output sound (C) at the mouth. The gain of the vocal tract (B) is thus measured only at integral multiples of the pitch frequency f_0. For the male voice we have chosen f_0 = 100 Hz (A1) and for the soprano female voice f_0 = 300 Hz (A2). In the idealized situations shown here, it is apparent that the resonances R1 and R2 can be estimated from the output sound if f_0 is low (C1). Determination of the resonances is much more difficult when f_0 is higher (C2). In contrast, the technique reported in this paper can measure the vocal tract response at intervals of 5 Hz or less. It is then possible to determine the resonances accurately irrespective of the pitch frequency.

One technique commonly used in speech coding and recognition to estimate the formants automatically is linear prediction (Makhoul 1975), which models the spectral envelope of the speech signal as an all-pole filter. The poles (or, more crudely, peaks in the magnitude spectrum) with sufficiently narrow bandwidth are then the estimates of the formant frequencies of the input speech signal.

2.2. Estimation from formants in whispered speech

During whispered speech, the source spectrum is produced by turbulent air flow through partially closed vocal folds and consequently contains many more frequency components than does normal speech. Whispered speech could thus allow more accurate determination of vocal tract resonance frequencies. The precision in real time is limited, however, because the source spectrum is unknown and fluctuating. Measurements thus generally require time-averaging for satisfactory results (Dowd 1995, Pham Thi Ngoc 1995). It is also possible that the articulation used during whispering differs from that used for normal voiced speech (Kallail and Emanuel 1984).

2.3. Estimation using an external source at the glottis

One means of overcoming the lack of resonance information present in the envelope of the speech spectrum is to excite the vocal tract artificially with a known signal and then to measure the response. Fujimura and Lindqvist (1971) used sinusoids, whereas Castelli and Badin (1988) used white noise, to excite the vocal tract externally at the level of the vocal folds. They then recorded the response at the lips. Hardware limitations required their subjects to sustain articulatory positions for an impractical length of time.

Djeradi *et al* (1991) used pseudo-random excitation of the vocal tract via the glottis. Since the response to this signal is uncorrelated with speech, this method allows speech production during measurement. Major limitations of these glottal excitation methods include the following.

(i) Accurate measurements require the excitator power to be three to four times that of the speech and this is uncomfortable for subjects (Djeradi *et al* 1991).

(ii) Measurements necessarily include the unknown transfer function of the cartilage and skin around the neck (Pham Thi Ngoc and Badin 1994).

2.4. Estimation using an external source at the lips

An alternative approach to glottal excitation involves exciting the vocal tract via acoustic coupling at the lips. A suitable real-time acoustic impedance spectrometer has already been developed by Wolfe *et al* (1994, 1995) for musical instruments. In this approach the spectrometer couples a broadband source of acoustic velocity flow to the vocal tract at the lips and measures the response at the lips using a microphone as a pressure transducer (Dowd *et al* 1996a,b).

In order to understand the response measured using this system, we must look at the frequency dependence of the vocal tract's impedance Z_{VT}, measured at the lips and in parallel with the radiant field. The external radiation impedance Z_E is given by

$$Z_E = \alpha z \frac{\mathrm{j}kr}{1 + \mathrm{j}kr} \tag{1}$$

where k denotes the wavenumber given by $k = 2\pi f/c$, $\mathrm{j} = \sqrt{-1}$, r denotes the radial distance, f denotes the frequency, c denotes the speed of sound, α denotes a geometrical factor determined by the solid angle available for radiation and z denotes the specific acoustic impedance (Fletcher and Rossing 1991). In the experiments reported here $f \leq 2.1$ kHz and the source, microphone and mouth opening are separated by no more than several millimetres. Consequently $kr \ll 1$ and then equation (1) simplifies to $Z_E \approx \mathrm{j}kr\alpha z$. The radiation impedance is thus almost entirely imaginary.

In the absence of the spectrometer, the tract will be loaded with the radiation impedance Z_E and it has resonances when the impedance at the lips is completely resistive, namely when $\mathrm{Im}(Z_{VT}) = -\mathrm{Im}(Z_E)$.

The spectrometer source is located near to the lips and we approximate it as a current source which drives the impedance $Z_\parallel$ produced by Z_{VT} and Z_E in parallel; that is

$$Z_\parallel = \frac{1}{1/Z_{VT} + 1/Z_E} \tag{2}$$

$Z_\parallel$ will therefore exhibit maxima at the resonances of the radiation-loaded vocal tract. Z_E increases only linearly with frequency whereas Z_{VT} has relatively sharp resonances. Consequently minima in $Z_\parallel$ will occur when minima occur in Z_{VT}, namely at the resonances of the unloaded vocal tract. To the extent that the vocal tract may be treated as a tube with a small radiation load at the lips, $Z_\parallel$ goes through a maximum at the resonance of the (loaded) vocal tract and then falls rapidly to a minimum. It is also possible to explain the form of the resonance in terms of an electrical analogy: the inductive radiation load appears in parallel with an impedance (the vocal tract) whose reactance changes sign abruptly at its resonance.

The positioning of the microphone and source close to but outside the lips is the result of an empirical compromise. If it is located too far from the lips, the spectrometer will measure the impedance of the radiation field, with relatively little contribution from the impedance of the vocal tract. A location inside the lips would allow measurement of an impedance dominated by that of the vocal tract, but is impractical for normal speech and the study of vowels with a small mouth opening or vowels that are associated with consonants.

A previous application of acoustic impedance spectrometry to the vocal tract (Dowd *et al* 1996a,b) required subjects to raise their soft palates in order to mime vowel sounds. Since the soft palate is unconsciously raised during voiced speech, this is a difficult movement to learn and introduces potential artefacts. The technique which we report in this paper retains the advantages of the system of Dowd *et al*, whilst allowing normal phonation and removing the disadvantage of requiring conscious palate movement. It has further advantages in resonance detection and real-time display.

3. The real-time acoustic resonance meter

3.1. The hardware

The acoustic excitation and calibration procedure was similar to that described by Wolfe *et al* (1994, 1995) and

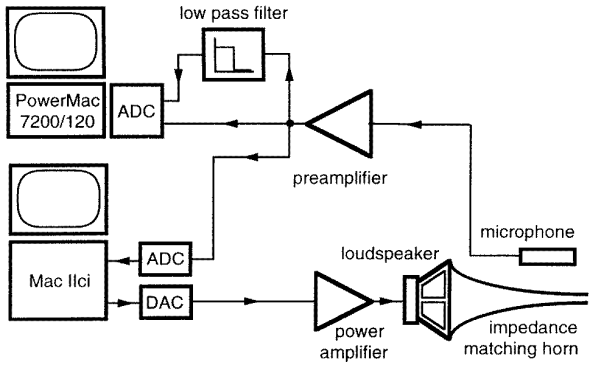

Figure 2. A schematic diagram of the apparatus used for real-time acoustic vocal tract excitation (RAVE) during phonation.

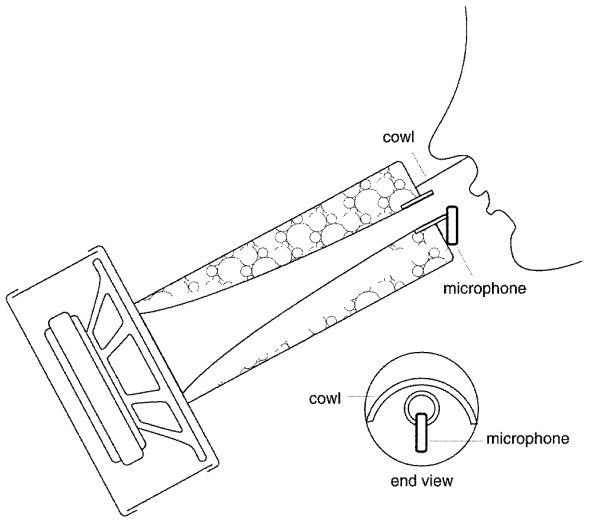

Figure 3. A schematic diagram (not to scale) indicating the construction of the acoustic current source and the configuration used to measure the acoustic impedance of the vocal tract. The exponential horn was cast inside a 380 mm length of PVC tube (65 mm outside diameter) and had a cut-off frequency around 160 Hz. The large end of the horn was driven by a 150 mm loudspeaker with a carbon fibre cone (JAYCAR RE/SPONSE RW6) mounted in a sealed enclosure packed with acoustic fibre. The small end of the exponential horn had a 16 mm internal diameter. The hemi-cylindrical cowl was 30 mm in length with a diameter of 65 mm. The sound pressure was measured by a small electret microphone (8 mm diameter, Tandy 33-1052).

used the circuitry and hardware shown in figures 2 and 3 respectively. A hemi-cylindrical cowl was attached to the upper surface of the end of the broadband excitation source to increase the acoustic coupling of the source to the subject's vocal tract, while leaving the lower jaw free to move. This cowl helped to maintain a constant distance from the upper lip to the excitation source and had the further advantage of reducing somewhat the distraction of the subject by the excitation source (the resultant sound level was always less than 75 dBA at the subject's ears).

The broadband source V_S was synthesized as the sum of k harmonic components of a fundamental frequency f_S via the equation

$$V_S(t) = \sum_{m=k_L}^{k_H} A_{Sm} \sin(2\pi m f_S t - \phi_{Sm}) \qquad (3)$$

where A_{Sm} denote the amplitudes and $k = k_H - k_L + 1$. The phases ϕ_{Sm} were selected at random to find a combination that significantly improved the signal-to-noise ratio (SNR) of the transfer function under measurement (Smith 1995). This excitation signal was generated by a computer (Macintosh IIci) via a 16-bit digital-to-analogue converter (DAC) (National Instruments NB-A2100). For calibration, the subject closed his or her mouth so that the reference acoustic load was the laboratory field with the lower face of the subject and the cowl as baffles. The spectrum of the microphone signal was calculated for this configuration and the amplitudes A_{Sm} adjusted so that the new measured spectrum for the signal applied to the reference load was frequency independent (flat) to within 2 dB.

Once it had been calibrated, the source remained unchanged for any series of measurements. When the subject opened his or her mouth, the vocal tract appeared acoustically in parallel with the free field baffled by the face and cowl, so that resonances of the tract appeared as strong variations with the frequency of the measured spectrum, as described above. The source was an acoustic current which was essentially independent of the load; consequently, in the absence of a speech signal, the spectrum of the measured pressure response was given by γ, the ratio of the acoustic impedance with the mouth open to the acoustic impedance with the mouth closed (the reference impedance), where

$$\gamma = \frac{Z_\parallel}{Z_E} = \frac{Z_{VT}}{Z_{VT} + Z_E}. \qquad (4)$$

Z_E depended only linearly on the frequency, so γ had maxima when $Z_\parallel$ had maxima, namely at the resonant frequencies of the vocal tract when it was loaded with the external radiation field.

During measurements, the pressure spectrum included both the signal due to the external excitation and that due to the subject's voice. This signal was analysed by the second computer (Power Macintosh 7200/120). For the results presented in this paper, the excitation source was configured to produce 354 sinusoids between the frequencies $f_L = k_L f_\Delta \approx 200$ to $f_H = k_H f_\Delta \approx 2100$ Hz, at a spacing of $f_\Delta = f_S = 5.383$ Hz. The frequency range 200–2100 Hz was chosen because we were only interested in the two lowest resonances for the particular application presented herein and restricting the frequency range improves the SNR. However, the current hardware allowed the frequency range to be extended up to 22 kHz if desired.

3.2. The measurement of the acoustic pressure signal

The excitation signal $V_S(t)$ will produce a response $V_R(t)$ at the lips of the form

$$V_R(t) = \sum_{m=k_L}^{k_H} A_{Rm} \sin(2\pi m f_S t - \phi_{Rm}) \qquad (5)$$

where A_{Rm} and ϕ_{Rm} denote the amplitudes and phases respectively.

The speech signal consisted of h harmonics (of significant magnitude) of a fundamental frequency f_V and had the form

$$V_V(t) = \sum_{n=1}^{h} A_{Vn} \sin(2\pi n f_V t - \phi_{Vn}) \qquad (6)$$

where A_{Vn} and ϕ_{Vn} denote the amplitudes and phases respectively of the nth harmonic. The pressure signal measured by a microphone just outside the mouth consisted of $V_R(t) + V_V(t)$. The output of the microphone was connected via a pre-amplifier (gain typically set at 38) to one channel of the stereo 16-bit analogue-to-digital convertor (ADC) of the Power Macintosh. The output of the pre-amplifier was also connected via a filter to the other channel of the ADC. Blocks of $N = 4096$ stereo pairs were sampled at $f_{SAMP} = 22.050$ kHz.

3.3. Pitch detection

The signal used for pitch detection passed through a low-pass filter (LPF) (a fourth-order Chebyshev switched capacitor with 2 dB passband ripple) connected to the other input channel of the ADC. This LPF was adjusted to the pitch of each individual speaker by observing its output spectrum and varying the filter's cut-off frequency (f_C) to maximize the ratio of the fundamental to the second harmonic. f_C was then kept constant for measurements of that particular speaker, since the pitch did not vary over a large range during our tests. (For higher pitched voices, for which $f_C > f_L$, the LPF was replaced by a bandpass filter). The fundamental component of the speech signal in this filtered signal exceeded other frequency components by at least 20 dB. A running estimate of the pitch frequency was then calculated from the time intervals between successive positive-going zero crossings.

3.4. Fourier transforms and spectral estimation

Our spectral estimation employed an $N = 4096$-point FFT. Since $f_{SAMP} = 22.050$ kHz, the harmonic spacing was $\Delta f = f_{SAMP}/N = 5.383$ Hz $= f_\Delta$. Spectral leakage will not occur when the response to the excitation signal is transformed because an integral number of cycles were always sampled at each frequency. Spectral leakage could occur, however, when the speech signal was transformed and suppression of the speech harmonics will then remove additional information from the response to the excitation signal. For this reason spectral leakage was reduced by applying a cosine window to the initial and final 10% of each block of data.

3.5. Possible approaches to determination of resonance frequencies

Several approaches for the determination of the resonance frequencies from a speech-corrupted excitation response were investigated via simulation (Epps 1996). One approach involved calculating the largest differences in local averages using the magnitude spectrum. This was unsuccessful because the magnitude of the speech harmonics in the speech-corrupted excitation response exceeded the magnitude of the resonances. Phase-spectrum methods (including group delay) provided no advantage. Adaptive techniques, which could have effectively suppressed the interfering speech signal, were slow and converged unreliably in this situation. The fastest and most robust approaches involved initial suppression of the speech component in the measured pressure spectrum; see sections 3.6 and 3.7.

3.6. The suppression of the speech signal

The speech harmonics appeared as narrow-bandwidth disturbances in a spectrum which was otherwise relatively smooth, except at the resonance frequencies (see figure 4(a)). The speech component was suppressed by replacing spectral points within a specified range ($\pm$20 Hz) of each integral multiple of the pitch estimate (see section 3.3) by linear interpolation (see figure 4(b)). The relative amplitude of the speech harmonics above 1 kHz was sufficiently small for their suppression to be unnecessary. Once the speech harmonics had been suppressed, it was then possible to calculate the complex impedance of the vocal tract as a function of the frequency.

3.7. The estimation of the resonance frequencies of the vocal tract

Once the speech signal had been suppressed, the resonance frequencies of the tract could be determined by inspection of the magnitude spectrum produced by the response to the broadband signal. The algorithm must work in cases in which a speech harmonic coincides with the resonance, consequently several of the harmonics in the response to the excitation signal will have been suppressed together with the speech harmonic. It must also work reliably when the SNR is poor. For these reasons, an algorithm detecting the steepest negative slope performed poorly in simulation trials. The approach selected involved identifying the resonances using the largest calculated differences in local average magnitude.

One advantage of the RAVE technique is immediately apparent in figure 4(a). In this example the resonance frequency is midway between the adjacent pitch harmonics. Inspection of the speech signal alone would not have revealed the presence of this resonance.

To identify the lowest two resonance frequencies for the application presented herein, local averages were calculated within $\pm$25 Hz for R1 over the frequency range 200–800 Hz and within $\pm$100 Hz for R2 over the frequency range 0.8–2 kHz (the ranges of R1 and R2 are known not to overlap for English and several other languages studied). These frequency ranges for the resonances were further refined to confine estimates to appropriate regions of the vowel plane for Australian vowel sounds (see figure 5); however, such adjustments must generally be tailored to the language or vowel set of interest.

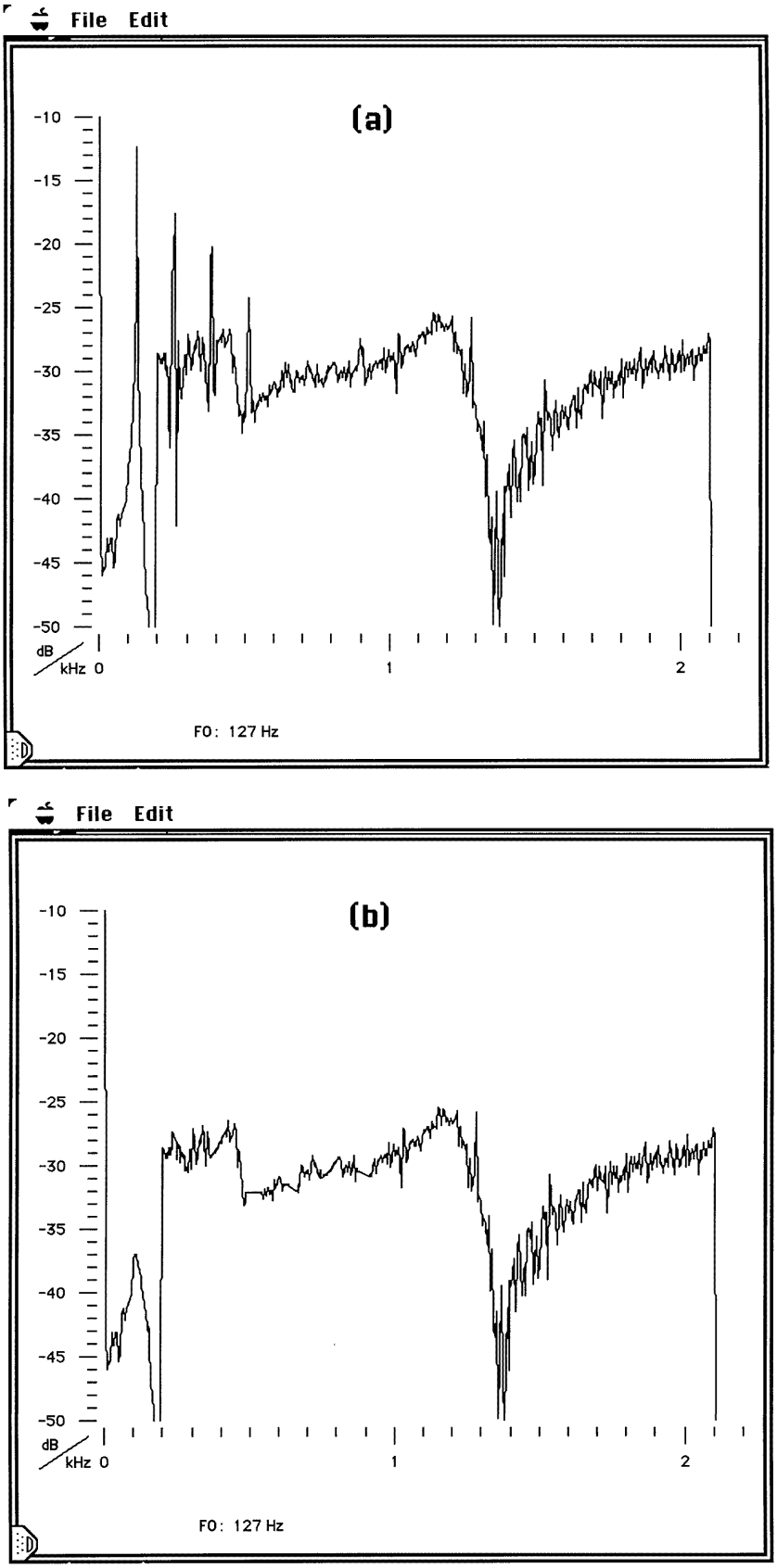

Figure 4. (a) The pressure signal magnitude spectrum showing the combination of speech signal and response to the broadband excitation signal. (b) The adjusted spectrum after suppression of the components of the speech signal below 1 kHz. The fundamental frequency of the speech signal was 127 Hz. The estimated resonance frequencies are apparent as maxima followed by sudden decreases in magnitude around 450 and 1300 Hz.

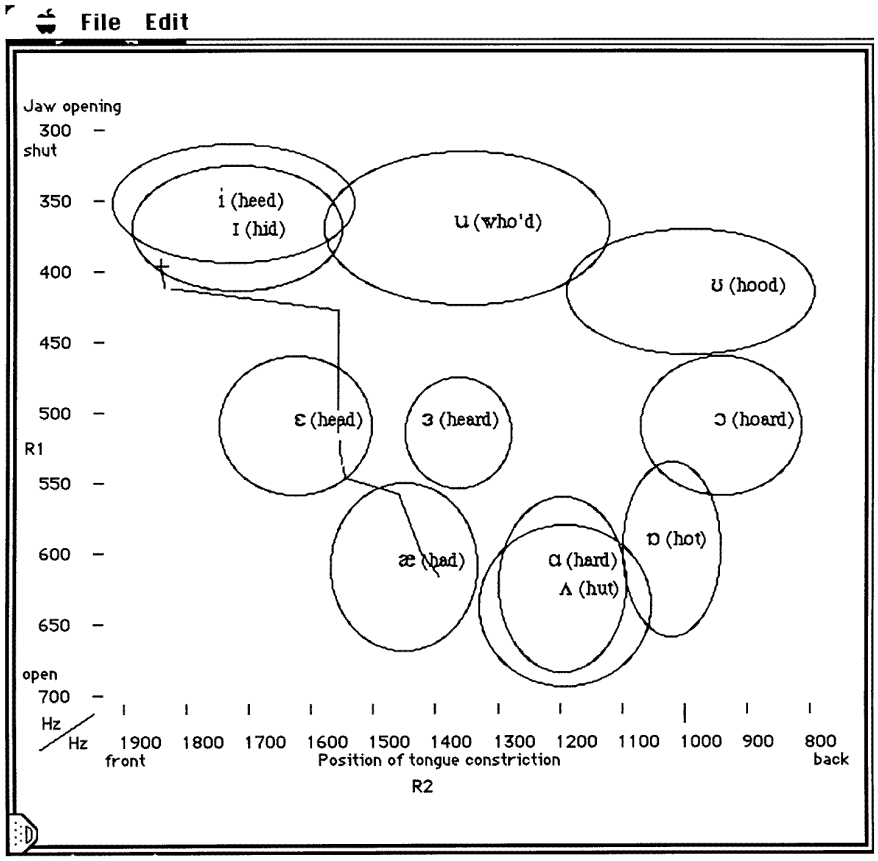

Figure 5. A resonance-plane plot showing the (R2,R1) trajectory of an Australian man pronouncing the word 'day', with the vowel regions for 33 Australian male subjects indicated by ellipses. The near coincidence of i and I and of ɑ and ʌ requires comment. In Australian speech, as well as in many other English dialects, these pairs are distinguished primarily by duration, ɑ and i being longer than ʌ and I .

3.8. The presentation of resonance estimates

The format in which the resonances are presented will depend upon the purpose of the measurement. If the resonances are to be used for speech therapy or language training, a useful format involves plotting the first two formants, F1 versus F2, with reversed frequency axes. This results in a vowel map similar to the map of vowel quality traditionally used by phoneticists (Clark and Yallop 1990). Vowel 'fronting' (the 'forwardness' of the tongue constriction in the mouth) has a large effect on the value of F2 and a smaller effect on F1. Vowel height (jaw opening) primarily affects the value of F1. Thus, for phonetic purposes, a useful representation of the resonance estimates involves plotting the point (R2,R1) in a plane with reversed axes (such as figure 5). The pitch frequency and amplitude of the fundamental are also displayed numerically. Use of only two formants might seem restrictive; however, we have demonstrated that such novel visual feedback can improve the pronunciation of vowels when it is presented as a spectrum (Dowd *et al* 1996a,b).

If three or more resonances are of interest, the magnitude and/or phase spectrum can be presented together with numerical values for the resonance frequencies. It is also possible for our new instrument to display information on the bandwidth or 'Q' of each resonance.

The instrument described essentially calculates the complex impedance of the vocal tract as a function of the frequency. RAVE can thus provide significantly more information than just the resonant frequencies, which substantially reduces the number of possible vocal tract configurations that could produce a particular sound. It could thus prove possible to calculate and to display the vocal tract configuration in real time (Sondhi and Resnick 1983, Schroeter and Sondhi 1994). This facility would provide a powerful research tool with considerable potential for language training and speech therapy.

4. The instrument's performance

4.1. The performance of the software

In order to achieve a display rate around 5 Hz, we set $N = 4096$ samples for all examples quoted here, which took an acquisition time of $t_{SAMP} = N/f_{SAMP} = 4096/22050 = 186$ ms. The harmonic spacing (Δf) was given by $\Delta f = 1/t_{SAMP} = f_{SAMP}/N = 5.383$ Hz. To eliminate problems with spectral leakage in the response to the excitation signal, the response was sampled over an integral number of cycles and this required that $f_S = \Delta f$. In the current version of the instrument, Δf was also the

upper limit to the display rate, which would be achieved by perfect data acquisition and infinitely fast processing.

Typical execution times for the individual components of each measurement cycle were: data acquisition 202 ms, pitch detection 1 ms, spectral estimation 14 ms, speech-harmonic suppression 0.5 ms, resonance detection 28 ms, graphical display 0.8 ms and total time 246 ms. The prototype instrument therefore measures and displays the resonance frequencies of the vocal tract at a rate slightly exceeding 4 Hz. This rate is sufficient for many interactive speech-training purposes. The display rate could be improved by the following measures.

(i) Reducing the time for data analysis. This could involve a faster processor and/or writing some routines in assembler. However, analysis currently occupies only 20% of the time between successive displays and so significant improvements in the display rate will be difficult and expensive.

(ii) Increasing Δf, thus decreasing the acquisition time at the expense of frequency resolution.

(iii) Utilizing overlapping sample frames with concurrent sampling and processing. It is then possible for the display rate to exceed Δf significantly.

Finer temporal resolution could also be achieved by recording the data and relaxing the condition that the resonances be displayed in real time. The technique of reducing the acquisition time by sampling for exactly one half of a complete cycle of the fundamental frequency offers no advantage in this situation because only the odd-order harmonics can then be used, which effectively halves the frequency resolution (Smith 1996).

4.2. The accuracy of pitch detection

The accuracy of the zero-crossing pitch-detection algorithm will be reduced by the presence of other frequency components in the filtered signal. This was tested by replacing the voice signal by a square wave of known frequency and similar amplitude. We found that the pitch estimate had a standard deviation of at most 1 Hz from the correct value, provided that the LPF cut-off frequency f_C was set within $f_V < f_C < 1.5 f_V$. When the bandpass filter was required, the pitch estimate had a standard deviation of at most 3 Hz from the correct frequency, provided that $f_V - 15$ Hz $< f_B < f_V + 50$ Hz (where f_B is the centre frequency of the bandpass filter).

In our measurements, subjects were instructed to speak at a level at which they could comfortably hear themselves over the excitation. They generally spoke at a peak level that was 10–20 dB above the average level of the excitation signal. The pitch frequency should thus always be determined to within a few hertz.

4.3. The accuracy of resonance determination

To test the accuracy of the speech-suppression and resonance-estimation algorithms, as well as any acoustic effects of the cowl on resonance frequencies, the vocal tract was replaced with an acoustic load with known resonance frequencies. This load had similar acoustic properties to the vocal tract during production of the sound *a* (as in 'hard'). It was made from one cylinder (30 mm diameter, 80 mm length) connected axially to a second cylinder (11 mm diameter, 90 mm length) closed at the other end (after Fant (1960)). The calculated resonances were confirmed independently of RAVE by measurements with a swept-frequency acoustic signal. A 'speech' signal was then simulated by electrically adding a square wave to the broadband excitation at the loudspeaker's input.

The standard deviations in the resonance frequencies measured by RAVE in the presence of this interfering 'speech' signal were then calculated. For soft 'speech' (peak power approximately equal to the average broadband excitation power) the standard deviation of the R1 estimate was typically 11 Hz about the value measured directly. This degraded to a 30 Hz standard deviation in R1 when the peak 'speech' power was increased to 20 dB above the average broadband excitation power. The estimate of R2 typically was within 3 Hz of the value measured directly and was essentially independent of the relative levels of speech signal and broadband excitation. Larger variations were observed in the estimate of R1 than in that of R2 during actual experiments due to the stronger speech harmonics at lower frequencies and a generally weaker R1 characteristic (see figure 4(a)).

4.4. A comparison with linear prediction

The sensitivity and robustness of the resonance estimates provided by RAVE were compared with those provided by linear prediction (LP) for the following Australian vowel sounds; ɛ (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'), i (as in 'heed'), and ɪ (as in 'hid'). 20 measurements of a sustained version of each vowel sound were taken from speakers with fundamental frequency 110 Hz (male, aged 22 years) and 205 Hz (female, aged 27 years) using both RAVE and a 24th-order real-time linear predictor applied to input data blocks of 25 ms duration.

The results of these measurements are shown in figure 6, in which the centre of each elliptical region is the point ($\overline{R2}$, $\overline{R1}$) while the semi-axes indicate the standard deviations. The variations presumably arose from variations in the vocal tract itself during measurement as well as experimental limitations.

The larger variation observed using linear prediction is consistent with the accuracy of this method being limited by the harmonic spacing of the speech signal. The performance of the linear prediction approach might be improved by more sophisticated variants, for example pole-focusing (Duncan and Jack 1988) or a complex variable-based approach (Snell and Milinazzo 1993), that were not employed in this study. Nevertheless, if the vocal tract response is sampled at Δf, then no method of interpolation, however sophisticated, can achieve a typical precision very much better than Δf. RAVE, on the other hand, exhibits much smaller variations in the resonance estimate, as a consequence of its much smaller harmonic spacing. The

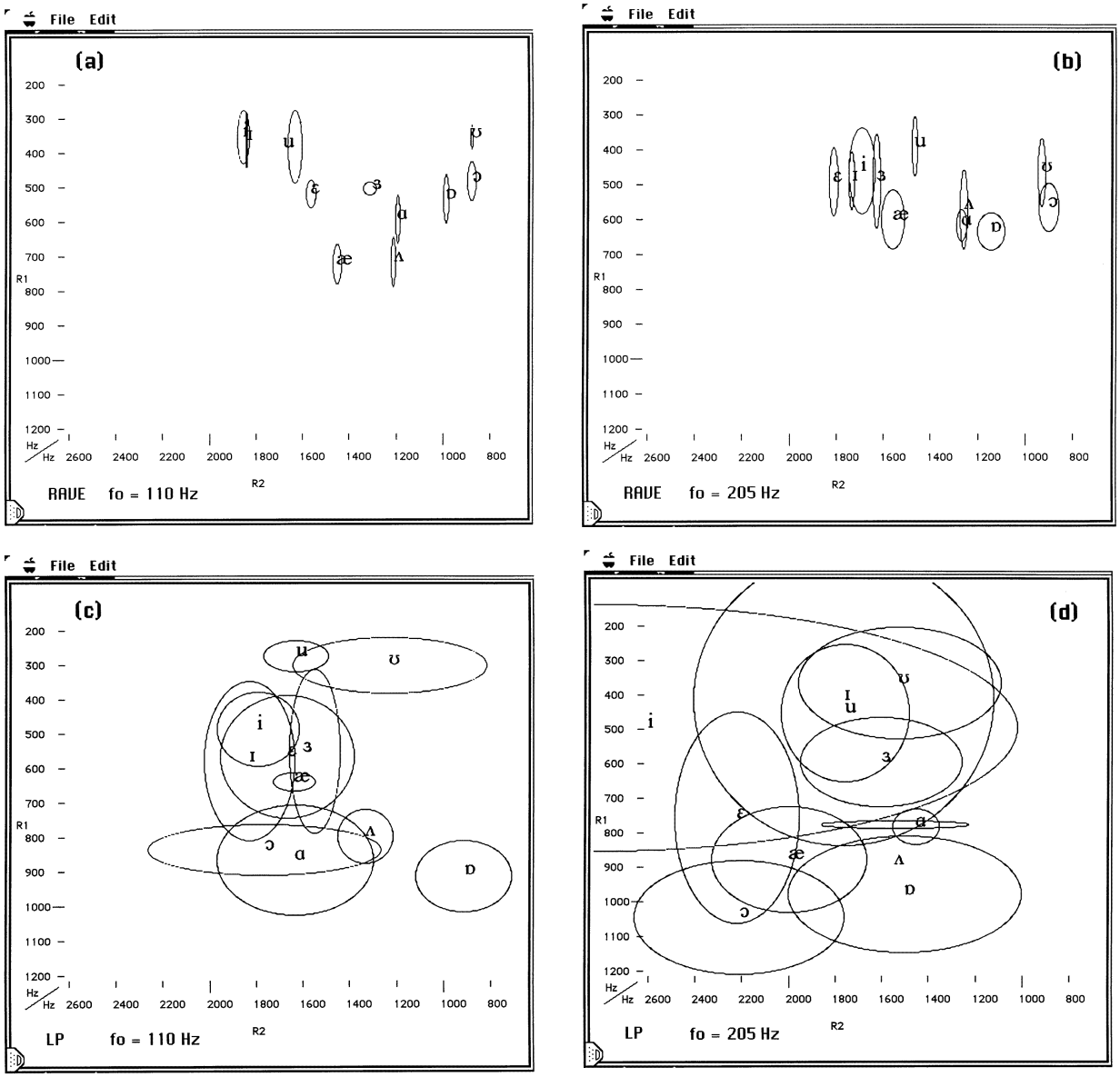

Figure 6. Resonance-plane plots of the position and variation in resonance estimates for real-time acoustic vocal tract excitation (RAVE) and linear prediction (LP). Two different speakers with pitch frequencies f_0 = 110 Hz and f_0 = 205 Hz were used.

variations in resonance estimates for RAVE also do not depend systematically upon the pitch frequency, unlike linear prediction, in which the variations increase with pitch frequency. Furthermore, the much smaller variation in results while using RAVE suggests that much of the variation in linear prediction estimates does not arise from sample variation.

4.5. Speech training

As mentioned previously, a desirable format for the visual feedback of articulatory parameters is a plot of the first versus the second resonance frequencies, with reversed axes, whereby the configuration being articulated at any time is represented by a point in this resonance plane. Such a plot is seen in figure 5, in which the 12 elliptical regions represent the standard deviations in each R1 and R2 about the point ($\overline{R2}$, $\overline{R1}$) for the vowel sounds of section 4.4 spoken by 33 Australian men (physics students at the University of New South Wales, Epps (1996)). Thus, the real-time cursor in this particular plane could be observed by a man learning the Australian vowel sounds and the difference between the cursor coordinate and the target region could be used to improve incorrect pronunciation. Similar target regions could be constructed for other languages to be learnt, by undertaking further such surveys.

5. Discussion

The technique of external vocal tract excitation presented here provides a non-invasive and accurate method of

measuring vocal tract resonances in real time for speaking subjects with an accuracy of the order of 10 Hz. The accuracy of resonance estimation by the processing of raw speech is limited by the fundamental frequency of the subject which is typically of order 100 Hz for men, 200 Hz for women and around 300 Hz for children. Thus, RAVE produces a significant improvement in the accuracy of resonance estimation for speakers with high-pitched voices over speech analysis techniques such as linear prediction.

The technique has obvious applications in acoustical phonetics because it permits direct, non-invasive measurement of the vocal tract resonances during phonation. It also has potential applications in speech therapy and language learning. Speakers with impaired hearing have difficulty in learning correct pronunciation of vowels because they lack auditory feedback and so the articulatory positions of the tract, especially those features inside the mouth and throat, are difficult to learn. Adults with normal hearing often have difficulty in learning accurate pronunciation of foreign languages. In this case the auditory feedback is complicated by the phenomena of categorization and interference (Landercy and Renard 1977, Clark and Yallop 1990): such learners hear a foreign phoneme but perceive it as a variant of a sound from their native language. They then produce an imitation which is closer to a sound already in their repertoire. Real-time display of the first two resonances provides useful feedback for accurate pronunciation (Dowd *et al* 1996a, b), presumably because it is not subject to categorization and because the resonance frequencies can be controlled by the subject by changing the jaw position and the tongue position.

Studies by Dowd *et al* (1996a,b) using an acoustic impedance spectrometer showed that subjects could use impedance spectra to imitate vowels with a success rate comparable to or better than that obtained using auditory feedback (with which the subjects had previously been familiar). These studies required conscious control of the palate while the speakers 'mimed' speech. We expect that the relaxation of this artificial constraint, using our new instrument, would give improved performance in this language-training application and also in speech-therapy applications.

Acknowledgments

The support of the Australian Research Council is gratefully acknowledged. Thanks are also due to Professor Neville Fletcher, Dr C Phillips and the volunteer subjects.

References

Castelli E and Badin P 1988 Vocal tract transfer functions with white noise excitation—application to the naso-pharyngeal tract *Proc. 7th FASE Symp. (Edinburgh)* pp 415–22

Clark J and Yallop C 1990 *An Introduction to Phonetics and Phonology* (Oxford: Blackwell)

Djeradi A, Guérin B, Badin P and Perrier P 1991 Measurement of the acoustic transfer function of the vocal tract: a fast and accurate method *J. Phonetics* **19** 387–95

Dowd A 1995 Real time non-invasive measurements of vocal tract impedance spectra and applications to speech training *Undergraduate Thesis, Medical Physics* UNSW

Dowd A, Smith J and Wolfe J 1996a Transfer functions of the vocal tract can provide real time feedback for the pronunciation of vowels *Proc. Australian Acoustical Society Conf., Brisbane* (Sydney: Australian Acoustical Society) pp 247–53

——1996b Real time, non-invasive measurements of vocal tract resonances: application to speech training *Acoustics Australia* **24** 53–60

Duncan G and Jack M A 1988 Formant estimation algorithm based on pole focusing offering improved noise tolerance and feature resolution *IEE Proc.* F **135** 18–32

Epps J 1996 Vocal tract excitation for real time formant estimation and speech training *Undergraduate BE Thesis* UNSW

Fant G 1960 *Acoustic Theory of Speech Production* (Gravenhage, The Netherlands: Mouton & Co)

Fletcher N H 1992 *Acoustic Systems in Biology* (New York: Oxford University Press)

Fletcher N H and Rossing T D 1991 *The Physics of Musical Instruments* (New York: Springer)

Fujimura O and Lindqvist J 1971 Sweep-tone measurements of vocal tract characteristics *J. Acoust. Soc. Am.* **49** 541–57

Kallail K J and Emanuel F W 1984 An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects *J. Phonetics* **12** 175–86

Landercy A and Renard R 1977 *Éléments de Phonétique* (Bruxelles: Didier)

Makhoul J 1975 Linear prediction: a tutorial review *Proc. IEEE* **63** 561–79

Pham Thi Ngoc Y 1995 Caractérisation acoustique du conduit vocal: fonctions de transfert acoustiques et sources de bruit *Thèse de doctorat* Institut National Polytechnique de Grenoble

Pham Thi Ngoc Y and Badin P 1994 Vocal tract acoustic transfer function measurements: further developments and applications *J. Physique IV* C **5** 549–52

Schroeter J and Sondhi M M 1994 Techniques for estimating vocal-tract shapes from the speech signal *IEEE Trans. Speech Audio Processing* **2** 133–50

Smith J R 1995 Phasing of harmonic components to optimize measured signal-to-noise ratios of transfer functions *Meas. Sci. Technol.* **6** 1343–8

——1996 Rapid measurement of transfer functions using less than one complete cycle *Meas. Sci. Technol.* **7** 110–2

Snell R C and Milinazzo F 1993 Formant location from LPC analysis data *IEEE Trans. Speech Audio Processing* **1** 129–34

Sondhi M M and Resnick J R 1983 The inverse problem for the vocal tract: numerical methods, acoustical experiments, and speech synthesis *J. Acoust. Soc. Am.* **73** 985–1002

Sundberg J 1987 *The Science of the Singing Voice* (De Kalb, Illinois: Northern Illinois University Press)

Wolfe J, Smith J, Brielbeck G and Stocker F 1994 Real time measurement of acoustic transfer functions and acoustic impedance spectra *Proc. Australian Acoustical Society Conf., Canberra* (Sydney: Australian Acoustical Society) pp 66–72

——1995 A system for real time measurement of acoustic transfer functions *Acoustics Australia* **23** 19–20