Real time measurements of the vocal tract resonances during speech.

Julien Epps, Annette Dowd, John Smith and Joe Wolfe
School of Physics
The University of New South Wales
Sydney 2052 Australia
jepps@newt.phys.unsw.edu.au

Abstract

The formants of speech sounds are usually attributed to resonances of the vocal tract. Formant frequencies are usually estimated by inspection of spectrograms or by automated techniques such as linear prediction. In this paper we measure the frequencies of the first two resonances of the vocal tract directly, in real time, using acoustic impedance spectrometry. The vocal tract is excited by a carefully calibrated, broad band, acoustic current signal applied outside the lips while the subject is speaking. The sound pressure response is analysed to give the resonant frequencies. We compare this new method (Real-time Acoustic Vocal tract Excitation or RAVE) with linear prediction and we report the vocal tract resonances for eleven vowels of Australian English. We also report preliminary results of using feedback from vocal tract excitation as a speech trainer, and its effect on improving the pronunciation of foreign vowel sounds by monolingual anglophones.

Introduction

In the source and filter model of voiced speech, the resonances of the vocal tract (the filter) modify the harmonic-rich signal from the vocal folds (the source). If the pitch of the source signal is sufficiently low, these resonances give rise to formants[1]: frequency bands of increased power in the spectrum of the sound radiated from the lips.

Formants are commonly estimated by inspection of spectrograms, or by automated routines such as linear prediction (LP). There is an inherent limitation to the precision of such estimates: they necessarily include an error which cannot be very much less than the frequency at which the tract's response is sampled, i.e. the pitch frequency. This imprecision may not be an important problem for many applications, especially if the speaker has a low pitched voice with a fundamental pitch of say 100 Hz or less. It is more important for higher pitched voices. For the voices of children and some women, the pitch frequency may be so high that the formants are poorly defined (Clark and Yallop, 1990). This is particularly the case when the pitch frequency is of the order of, or greater than, the frequency of the first resonance (Sundberg, 1987).

The relatively poor precision in formant estimation obtained from excitation by the speech signal is usually not a problem when the signal is used for recognition by human listeners. Normal spoken and written languages have a high level of internal redundancy (Fletcher, 1992) and so, with appropriate semantic and linguistic context, high word recognition rates may be achieved with poor resolution of the vowel formants, or sometimes even in the [k-mpl-t -bs-ns -v v--lz]. If automated recognition systems are to identify phonemes successfully without contextual clues, greater precision in formant or resonance estimation would be useful.

Speech training is another application which requires precise measurements in real time. People with very poor or no hearing have difficulty learning accurate speech because they lack auditory feedback. Adults learning foreign languages rarely acquire good accents because their auditory feedback is complicated by categorisation and interference: they interpret a foreign phoneme in terms of one in their native language and then reproduce a sound more like that with which they are familiar. A real time system which measures accurately the relevant articulatory properties of the vocal tract can be used to give feedback which is not compromised by categorisation and interference (Dowd et al, 1996).

Other investigators have used various methods to determine vocal tract resonances with higher precision than that available from voiced speech excitation. Pham Thi Ngoc and Badin (1994) report measurements made by exciting the tract mechanically near the glottis. This technique is good for measuring the vocal tract transfer function, but is perhaps less suitable for application to speech training because it is a little invasive and because it is also possible that phonation during such stimulation is not the same as phonation in the absence of external mechanical excitation. The resonances may also be determined from the spectrum of whispered speech, but the signal is noisy and unpredictable, so time averaging is necessary for precise measurement (Pham Thi Ngoc, 1995; Dowd, 1995).

We describe a system for rapidly, precisely and non-invasively measuring the resonances of the vocal tract

[1] The term formant is sometimes used to describe both the resonances and the local maxima in the sound spectrum. We measure the two independently, so we reserve the terms "resonance" for the former and "formant" for the latter. We refer to them as (R1,R2,...) and (F1,F2,...) respectively.

during phonation, its application in the measurement of resonances for 11 vowels in Australian English, and the preliminary results of using feedback from vocal tract excitation as a speech trainer.

Materials and Methods

Impedance spectrometer. The measurements use a development of an acoustic impedance spectrometer described by Wolfe et al (1995); Wolfe and Smith (1995). A signal which comprises the sum of several hundred sine waves is synthesized by a computer, converted to an analogue signal, amplified and input to an enclosed loudspeaker connected to the large end of an exponential horn. Near the other end of the horn (the source) is a small microphone (Fig 1), and both are positioned just outside the subject's lips, with a hemi-cylindrical cowl touching the face just below the nose. The signal is calibrated with a reference acoustic load: in this case it is the free field near the subject's mouth, with the subject's face (mouth closed) and the cowl acting as baffles. During calibration, the amplitude of each of the sine waves is adjusted so that the resulting spectrum measured by the microphone with the subject's mouth closed is independent of frequency or "flat". When the subject opens his/her mouth to speak, the vocal tract is in parallel with the acoustic field and its resonances appear as maxima followed by steep falls in the broad band response, as shown in Figs 2c and 2d. To determine these from the total signal, the speech signal is removed. This is done by measuring its pitch (using a high order low pass or band pass filter and zero-crossings) and then removing integral multiples of this frequency (± 20 Hz) from the combined signal. These gaps are filled by interpolation. A routine searches for the largest negative going discontinuities between levels of adjacent frequency bands averaged over 25 Hz (for R1) and 100 Hz (for R2). The frequency spacing is about 5 Hz but, because of the calculations, each cycle takes a little longer than 200 ms and new values for R1, R2 and other data are displayed four times per second. Technical details and the performance of the method are described elsewhere (Epps et al, in press).

Australian vowels. The vocal tracts of 33 young Australian men (students at the University of New South Wales) were measured. The acoustic source and microphone were positioned outside their mouths, with the microphone about 10 mm and the source about 30 mm from their lips. An instruction sheet asked them to pronounce and to sustain the words "heed" [i], "hid" [ɪ], "head" [ɛ], "had" [æ], "hard" [ɑ], "hot" [ɒ], "hoard" [ɔ], "hood" [ʊ], "who'd" [u], "hut" [ʌ] and "heard" [ɜ]. 15 measurements were taken for each vowel.

Comparison with Linear Prediction. To compare RAVE and the standard automated technique of linear prediction (LP) (Makhoul 1975), we measured the formant frequencies of one male and one female subject. Linear prediction (LP) is a spectral smoothing process which fits a curve to the envelope of the speech signal and reports peaks in the fitted curve. Subjects were requested to produce the same set of 11 Australian vowels, in their usual conversational voice, into a microphone. A 24th order LP model was fitted to their speech in real time, and the formant frequencies were estimated by calculating the frequencies of the two lowest-frequency poles of the model. 20 estimates of their first two formant frequencies were recorded for each vowel sound. The same two speakers were also requested to produce the same vowels while their vocal tract resonances were measured using RAVE, as described above.

Speech trainer. The experiment using vocal tract feedback as a speech trainer used an earlier version of the spectrometer and software which did not allow phonation. For this study, subjects were taught to raise the soft palate (velum) and to mime the production of a vowel. "Target" sounds were chosen from the French language, as spoken by 5 female native speakers of the language. The sounds were recorded on magnetic tape, and the frequencies of the first two resonances of the tract were measured. 11 monolingual Australian women volunteers were asked to imitate these target sounds. One group received only auditory feedback - they listened to the sounds and attempted to reproduce them. Another group used both auditory feedback and vocal tract feedback. The latter consisted of displaying to the subject the acoustic response of her own vocal tract, upon which was superimposed the values of the resonant frequencies of the target speaker. Tape recordings of subjects attempting to imitate the target sounds were distributed to a listening panel of native speakers who indicated which vowel they thought that the subject had produced. The procedure is described in detail in a manuscript submitted elsewhere.

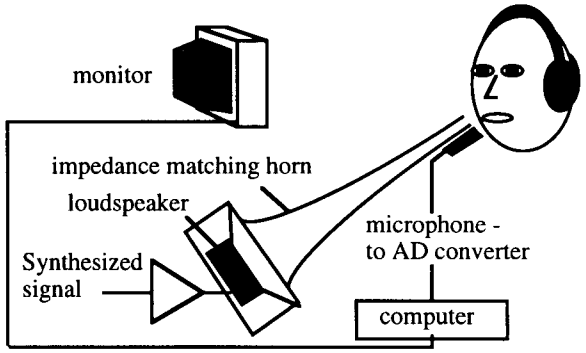

Fig. 1. A synthesized signal containing several hundred frequency components is input to the amplifier, speaker and impedance matching horn. The microphone signal includes both the subject's voice and the response of the vocal tract to the external signal. The monitor may be used to give the subject instructions for recording data, or to display visual feedback about his/her vocal tract.

Results

The operation of RAVE on a male and female voice is shown in examples in Fig. 2. (a) and (b) show the spectrum of the voices alone, while (c) and (d) show the spectrum measured with both phonation and external excitation. To compare RAVE with linear prediction (LP), 20 measurements were made of the resonances and formants for each of 11 vowels of Australian English as spoken by one male and one female speaker. The reason for single subject samples was to minimise intra-sample variation. The first two resonances were measured using RAVE. The first two formants were measured using LP. In each case RAVE gave a smaller variation (Table 1).

	Male Voice		Female Voice		
	σ(F1,R1)	σ(F2,R2)	σ(F1,R1)	σ(F2,R2)	
LP	122	240	206	498	Hz
RAVE	64	16	91	31	Hz

Table 1. The average standard deviations in the formants (using LP) and the vocal resonances (using RAVE) for 20 measurements each of 11 Australian vowels.

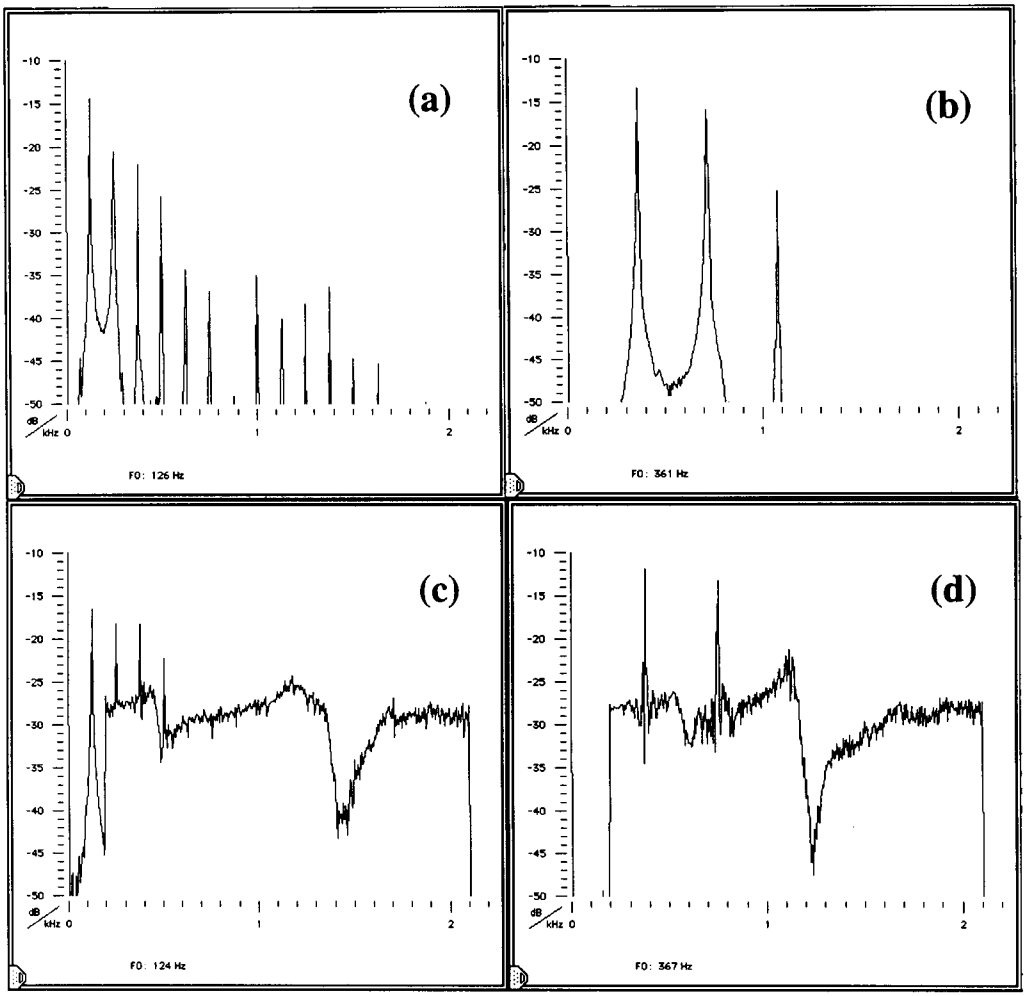

Fig. 2 (a) is the magnitude spectrum for a speaker with pitch frequency 126 Hz. From this, one can estimate formants at about 0.5 and 1.4 kHz. (b) is the spectrum for the same vowel spoken by a speaker with pitch frequency 361 Hz. From this spectrum it is much more difficult to estimate formant frequencies. The RAVE technique uses a calibrated broad band source to excite the vocal tract from just outside the lips. In this example it contains 354 sine waves with frequencies equally spaced over the range 0.2 to 2.1 kHz. (c) and (d) show the spectra measured with the vocal tract excited by the broad band source during pronunciation of the same vowels. At a modest conversational level, the harmonics of the speaker's voice are seen above the broad band spectrum at frequencies less than about 1.6 kHz. The maxima followed by large negative slopes in the broad band response at 0.4 and 1.3 kHz (in c) and at 0.5 and 1.1 kHz (in d) are due to the resonances of the vocal tracts in the configuration for these vowels.

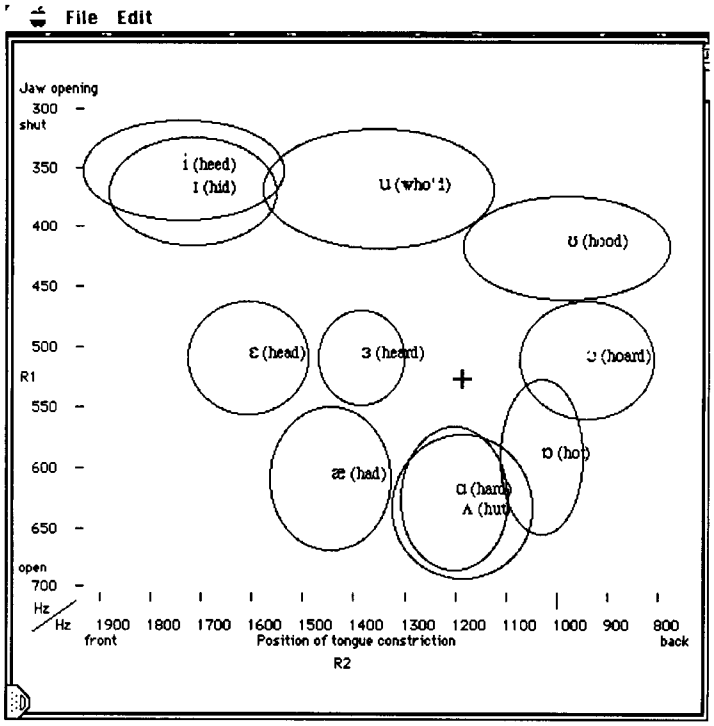

Fig. 3. Measurements of the vocal tract resonances (R2,R1) of 33 young Australian men pronouncing 11 Australian vowels. ($\overline{R2}$,$\overline{R1}$) is at the centre of each ellipse and the semi-axes are the standard deviations. The figure is a "screen dump" of a version of the RAVE technique used as a speech trainer, and for that reason only the axes are also labelled "Position of tongue constriction" and "Jaw opening". These parameters were not measured directly, but they are correlated with R2 and R1. A cursor on the monitor (the cross at 0.52, 1.20 kHz) shows the current configuration of the user's own vocal tract in real time, and the ellipses are used as targets which change colour when "hit" by the cursor.

Fig. 3 shows the first two resonances of 11 vowels of Australian English as measured on 33 Australian

men. This figure also displays the screen used as the application of RAVE to speech training. In this application, target areas are shown on the screen, along with the current configuration (R1,R2) of the user's vocal tract, shown as a moving cursor. The axes are labelled as jaw position and tongue position respectively as an initial aid to moving the cursor. After a little training, users can steer the cursor just by "thinking about where they want it to go": it is a little like a video game, but with mouth control rather than a joy stick.

Fig. 4 shows preliminary results from a study to investigate the use of feedback about the vocal tract as a speech trainer for foreign language teaching. Six vowels from French were chosen as targets. These may be considered as three pairs of vowels which are often confused by non-native speakers. /a/ and /ɑ/ are acoustically similar and are occasionally confused by native speakers. /e/ and /ɛ/ are also rather similar. /u/ and /y/ are quite different acoustically. They are virtually never confused by native speakers of French, but relatively often confused by English speakers.

Fig. 4 shows that subjects who used vocal tract feedback as well as auditory feedback produced values of R1 and R2 which were very similar to those of the target native speakers. Tape recordings of the sounds made by the different groups were played by a listening panel of native speakers who were asked to identify the vowel sounds. The recognition rate was significantly higher for the group using both types of feedback.

Conclusion

Auditory feedback is a model of standard language teaching in which students hear a sound and attempt to imitate it. This is a method that they have been using all their lives. Vocal tract feedback, on the other hand, is a new type of feedback and involves novel coordination between eye and articulation. Nevertheless, one to two hours training with this feedback significantly improved the articulation and comprehensibility of our subjects. These preliminary results (to be reported in detail elsewhere) suggest that RAVE has considerable potential in language laboratories and in speech pathology.

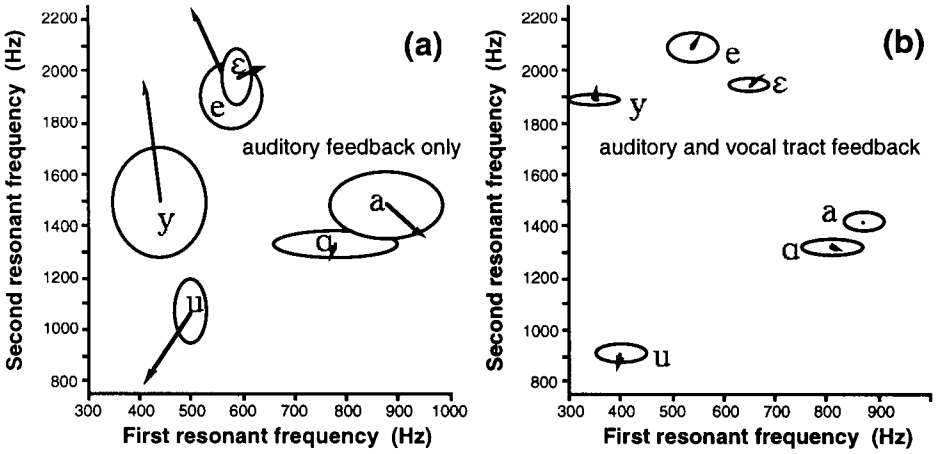

Fig. 4. Measurements of the first two vocal tract resonances by monolingual anglophone subjects attempting to produce six "target" vowels spoken by native French speakers. The head of each arrow is the target value, the tail is the average for the subjects and the semi-axes of the ellipses are the standard deviations among the subjects. (Short arrows are good imitations, small ellipses show little variability.) (a) shows the results for the subjects using auditory feedback only. (b) shows the results for subjects who had spent one hour learning vocal tract feedback and who were then given both auditory and vocal tract feedback to imitate the target vowels.

Acknowledgements. We acknowledge support from the Australian Research Council and thank our volunteer subjects.

Patent. The technology is the subject of provisional patents and the authors would welcome enquiries from companies interested in manufacturing the device.

References

Clark, J. and Yallop, C. An Introduction to Phonetics and Phonology, (Blackwell, Oxford 1990).

Dowd, A. Real time non-invasive measurements of vocal tract impedance spectra and applications to speech training. Undergraduate thesis, Medical Physics, UNSW Sydney (1995).

Dowd, A.; Smith, J. and Wolfe, J. Real time, non-invasive measurements of vocal tract resonances: application to speech training. Acoustics Australia 24: 53-60 (1996).

Epps, J., Smith, J.R. and Wolfe, J. "A novel instrument to measure acoustic resonances of the vocal tract during speech" Measurement Sciecne and Technology, in press.

Fletcher, N.H. Acoustic Systems in Biology (Oxford, NY 1992).

Makhoul, J. Linear Prediction: A Tutorial Review. Proc. IEEE 63: 561-579 (1975).

Pham Thi Ngoc, Y. Caractérisation acoustique du conduit vocal: fonctions de transfert acoustiques et sources de bruit" Doctoral thesis, Institut National Polytechnique de Grenoble (1995).

Pham Thi Ngoc, Y. and Badin, P. Vocal tract acoustic transfer function measurements: further developments and applications. J. de Physique IV C5: 549-552 (1994).

Sundberg, J. The Science of the Singing Voice, (Northern Illinois Univ. Press., De Kalb, Ill 1987).

Wolfe, J. and Smith, J. A comparison of acoustic impedances of flutes - a preliminary study. Intl. Symposium on Musical Acoustics, Dourdan, France. 100-106 (1995).

Wolfe, J., Smith, J.,Brielbeck. G., and Stocker, F. A system for real time measurement of acoustic transfer functions. Acoustics Australia 23: 19-20 (1995).