Acoustic cues of voiced and voiceless plosives for determining place of articulation

Philip J.B. Jackson

School of Electronic and Electrical Engineering, University of Birmingham,
Edgbaston, Birmingham B15 2TT, UK. Email: p.jackson@bham.ac.uk

Abstract

Speech signals from stop consonants with trailing vowels were analysed for cues consistent with their place of articulation. They were decomposed into periodic and aperiodic components by the pitch-scaled harmonic filter to improve the quality of the formant tracks, to which exponential trajectories were fitted to get robust formant loci at voice onset. Ensemble-average power spectra of the bursts exhibited dependence on place (and on vowel context for velar consonants), but not on voicing. By extrapolating the trajectories back to the release time, formant estimates were compared with spectral peaks, and connexions were made between these disparate acoustic cues.

1. Introduction

It is acknowledged that there are three principal cues to the place of articulation for voiced and voiceless plosive consonants: (i) the spectral characteristics of the burst, (ii) the formant trajectories leading into and out of the stop closure, and (iii) visual cues to the position of the tongue, lips, etc. [1]. Only the first two, the acoustic cues, are considered in this paper, which nevertheless aims to make some comparison between them through detailed analysis of speech signals from those mixed-source sounds. In English, the voiceless stops /p,t,k/ typically comprise the burst at release (plosion), followed by a short spell of frication and finally aspiration, before or during the onset of the succeeding phone [2]; the voiced stops /b,d,g/ also have a contribution from phonation, and that from aspiration tends to be lesser.

The speech signals were decomposed, using the pitch-scaled harmonic filter (PSHF), into separate periodic and aperiodic components corresponding to the contributions from voicing and turbulence noise, respectively [3]. The PSHF is a technique that has previously revealed details of turbulence noise in vowels and fricatives [3, 4], and is known to perform best during steady voicing [3, 5]. Hence, the formant frequencies of the components were estimated by linear prediction, implicitly assuming an all-pole model of the vocal-tract transfer function. But, although there are anti-resonances (or zeros) in the transfer function for turbulence-noise sounds owing to the location of the source, e.g., for the burst of plosion emitted at release, the resonances (or poles) are the same as for vocalisation [3]. So, if we can localise the resonances at release, i.e., the formants, we will have gained partial knowledge of the overall transfer function. As the formant values are reliably extracted only during the neighbouring regions of voicing, we need to define a model that can be used to project the formant tracks back to the instant of release. This will enable us to make a direct comparison with the peaks of the burst spectra (the first cited cue).

For many years, the formant trajectories following a stop consonant have been given linear (or static) descriptions [2, 6]. Analogous models have been applied in automatic speech recognition with a view to capturing aspects of the speech dynamics [7]. Recently, others have experimented with exponential or state-space models [8, 9], and inspection of a typical spectrogram would lead one to consider them a more natural choice. This is particularly true for the first hundred or so milliseconds after the release of a plosive, when formant transitions are not generally linear and tend to change most rapidly at onset. Following in this vein, we have fitted exponential functions to the formant tracks, which are later used to predict the formant frequencies at release. Moreover, the exponential trajectories provide a more robust estimate of the formant frequencies at voice onset, that is, of the loci — the conventional cue for determining a plosive consonant's place of articulation [2, 6, 10, 11]. Finally, we computed power spectra centred on each plosive burst, which were ensemble averaged to improve the quality of the measurements.

2. Method

High-quality speech recordings were made of an adult male speaker of British English R.P. (PJ) in an anechoic chamber, with a head-mounted microphone connected to the 16-bit soundcard of a PC via a pre-amplifier. The data were sampled at 16 kHz and transferred digitally for analysis. The corpus contained CV syllables, C = /p,t,k,b,d,g/ and V = /a,i/, with ten instances uttered per breath.

Optimisation of the pitch-period estimate was performed from a manual initialisation for each segment of voicing, which was then decomposed by the PSHF into periodic and aperiodic components (10-point frame offset, amplitude-based time-series outputs, $\hat{v}(n)$ and $\hat{u}(n)$). The results of each decomposition were respliced with the original signal to yield a complete pair of periodic and aperiodic signals.

Formants were extracted from the periodic component using the SFS software package, which estimates the formant frequencies from linear prediction coefficients (LPC, autocorrelation, 20 ms window, 12 poles) [12]. During periods of voicing, these formant estimates matched the formant bands from spectrograms, and were more consistent than those extracted from the original speech signal (or from the aperiodic component).

The formant transitions following release of a plosive have traditionally been described primarily in terms of the formant frequency at onset in relation to its asymptote in the steady-state vowel (i.e., at transition offset) [6, 11, 13]. Clearly, a more accurate estimate of the formant's locus would take account of the whole sequence of formant measurements during the transition. We chose to do this by modelling the formant trajectories

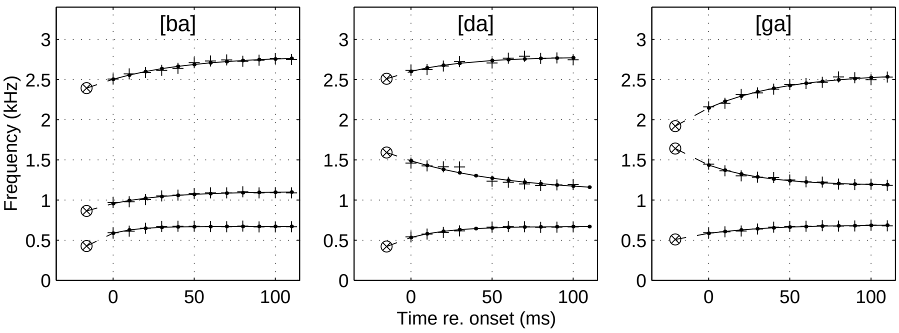

Figure 1: Typical examples of formant trajectories (•, solid) estimated from measurements every 10-ms frame (+), and extrapolated to give their loci at release (⊗, dashed), shown for: (left) [ba], (centre) [da], and (right) [ga].

Formant	const	linear	quad.	exp.
F1	31.0	19.5	8.2	7.5
F2	43.4	20.2	8.9	8.7
F3	65.9	23.3	16.4	16.8

Table 1: Root mean-squared error of the formant values for [ba] with respect to the optimally-estimated trajectories (Hz): constant, linear, quadratic, and exponential models.

with an exponential function:

$$x_t = c\left(1 - k^t\right) + x_0, \tag{1}$$

where x_0 is the trajectory's value at onset ($t = 0$), the asymptote is $c + x_0$ and k is the rate constant, whose magnitude is less than one for a decaying response. Assuming no prior knowledge of the parameters, we can find the optimal estimates by adopting a least-squares approach. For c and x_0, closed-form solutions can easily be derived:

$$\hat{c} = \frac{\sum_{t=1}^{T}\left(y_t - x_0\right)\left(1 - k^t\right)}{\sum_{t=1}^{T}\left(1 - k^t\right)^2}; \tag{2}$$

$$\hat{x}_0 = \frac{\sum_{t=1}^{T} y_t - c\sum_{t=1}^{T}\left(1 - k^t\right)}{\sum_{t=1}^{T} 1}. \tag{3}$$

(These formulae can be simplified for contiguous data, as geometric series.) Unfortunately, for k, the derivation does not trivially provide a closed-form solution, and so the optimal estimate $\hat{k}$ was sought iteratively.

3. Results

Formant tracks during sections of voicing were identified, and the first half of each section was used for investigating their form following release. Some of the formant data from the LPC analysis were spurious, largely owing to the selection of higher frequency peaks from the spectrum. These values were marked by hand and simply omitted from the trajectory calculation. This calculation yielded our first set of trajectory models, and is illustrated in Figure 1, for each place of articulation. The trajectories were then extrapolated back to the time of release (as determined by the first substantial spike in the speech signal after closure), also shown in Figure 1.

By comparison with parametric models of equal or fewer degrees of freedom, the exponential function did indeed perform best on average, as demonstrated by the RMS errors between the trajectories and the measured formants, listed in Table 1. However, such freedom occasionally produced exaggerated trajectories with too high curvature, often triggered by a single extraneous formant reading. In order to curb these artefacts, the rate constants were constrained by setting them to their average values (for each combination), and the trajectories re-estimated accordingly.

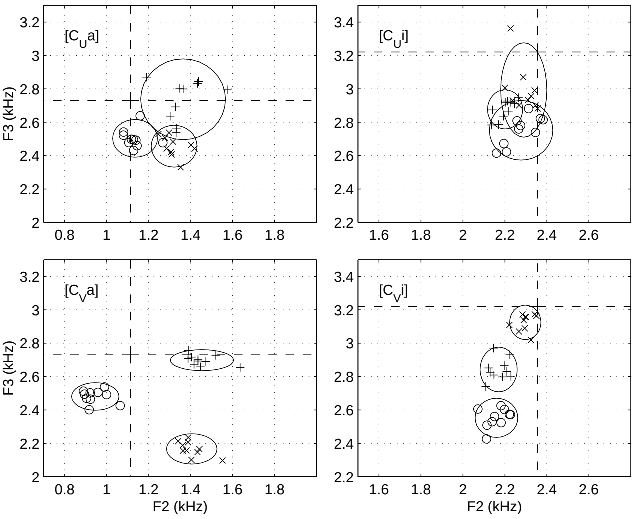

Figure 2: Scatter plot of estimated F2 versus F3 loci at voice onset after the plosive consonants C_U = /p,t,k/ (upper panels) and C_V = /b,d,g/ (lower panels) in the context /CV/, where V = /a/ (left), and /i/ (right): labial (○), alveolar (+), and velar (×). Ellipses representing the distributions are drawn using the sample means and variances with a radius of two standard deviations. The dashed grid lines mark the steady-state vowel formant frequencies.

3.1. Formant loci

As a means of characterising the place of articulation, the trajectory positions at release were examined. The predicted F2 and F3 values provided a strong cue for determining the place of the voiced plosives, much in line with previous studies that used the locus at voice onset [10, 13]. As expected, the results were more distinct for the open /a/ context than for /i/. However, the absence of voicing during most of the transition in

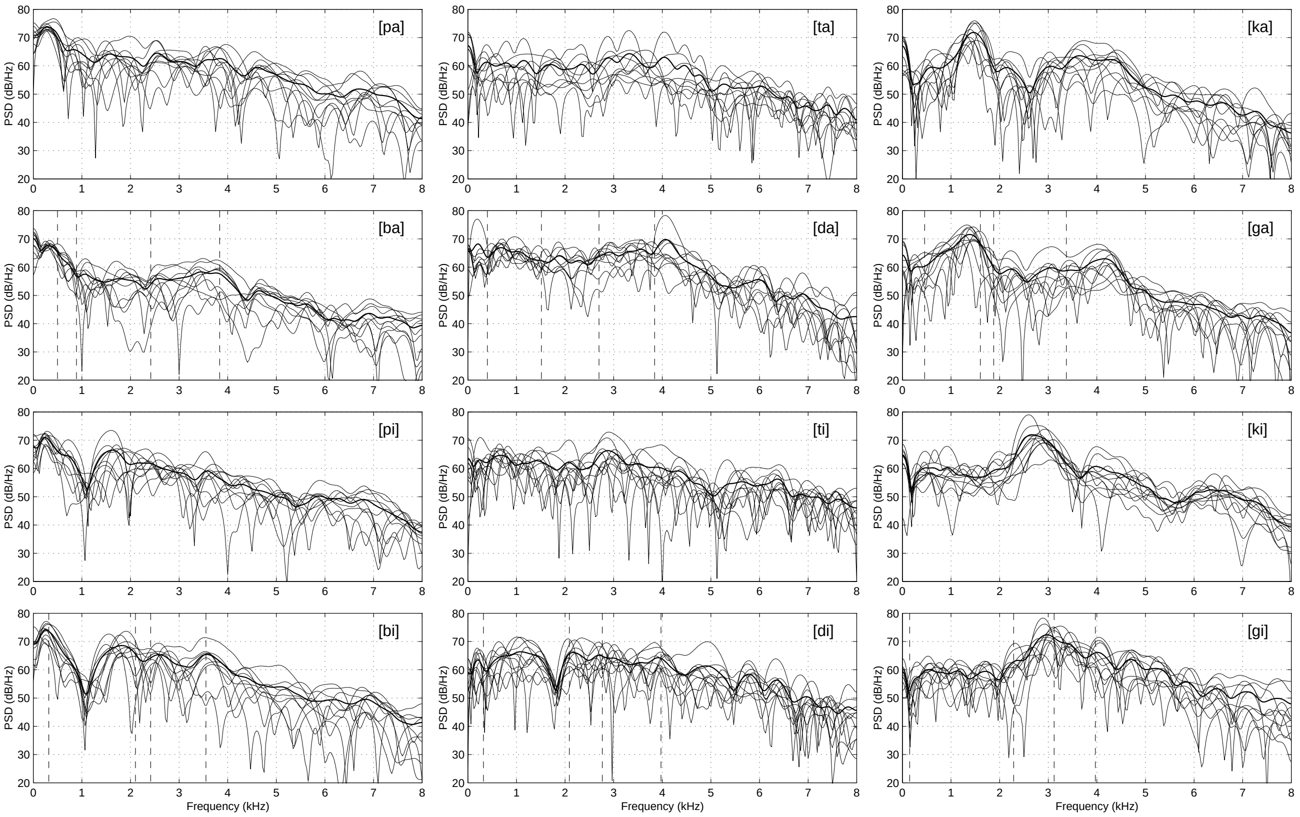

Figure 3: Power spectral densities (6 ms Hann window, ×4 zero padded) of plosives in /a/ context (upper half) and /i/ context (lower half), for each token (thin lines) and their ensemble average (thick line): (from left) labial, alveolar and velar. The upper row within each half shows voiceless plosives, [p,t,k], and the row below voiced, [b,d,g]. Also shown as vertical dashed lines are the mean predicted formant loci for F1, F2 and F3, and the mean formant frequency F4.

voiceless plosives resulted in wide variations in the predictions, although there was some (inconclusive) evidence to suggest that the underlying behaviour was not inconsistent with their respective voiced counterparts. Yet these features were found to be no better for discrimination than the standard loci. Those derived from the F2 and F3 trajectories at onset are plotted in Figure 2.

The loci are presented separately for voiced and voiceless stops and by vowel context, from which it is immediately apparent that, while the clusters for [b,d,g] are quite distinct, the voiceless ones are overlapping and nestle around the formant frequencies of the subsequent vowel. The trajectories approach the steady state from directions that are more different in the /a/ than the /i/ context, and therefore the [Ca] clusters are more distinct than those for [Ci]. Thus, with /i/'s higher F2, the clusters for [bi] and [di] become indistinguishable in the F2 direction. In the F3 direction, we see a dramatic change in the position of the locus for [g], from being below those of [b] and [d] to above them, which accords with previous reports of the effect of front- or back-vowel context on velar placement [10]. The distributions of loci are generally wider for the voiceless stops than for the voiced ones, whose variances tended to increase with distance of the locus from the steady-state vowel's formant frequencies.

3.2. Ensemble spectra

In a separate analysis, short windows (6 ms) were centred on each burst of plosive release (determined manually), zero-padded to 512 points (30 ms), and their power spectral densities (PSDs) averaged to give the ensemble spectrum for each CV combination. The individual spectra for each of the ten tokens in the ensemble, and their averages are drawn in Figure 3.

The most notable similarity is between the voiced/voiceless cognates, and is even more pronounced than between contexts for the same phoneme, for example, [b(a)] is more similar to [p(a)] than to [b(i)]. The effect is most striking for the velar stops, where the change in vowel context has implcations for the position of the tongue, and manifests itself as a dramatic modification of the spectral shape. The predominant peak shifts from c. 1.4 kHz to approximately double. It is also noteworthy that spectral troughs occur repeatedly for certain combinations, heralding the presence of vocal-tract anti-resonances for the plosive source, such as at 1.1 kHz for [pi,bi], c. 1.8 kHz for [di], and c. 2.7 kHz for [ka]. However, the main auditory effect of the bursts comes from the frequencies of high spectral energy.

As F2 and higher formants are practically cancelled by zeros due to the source location at the lips, for labial stops, their spectra are relatively flat and dominated by the low-F1, Helmholtz resonance. The alveolar stops have a broad high-frequency peak, rather like the fricative /s/, largely generated by acoustic dipoles at the incisors that excite F4 and F5. To understand why the strong mid-frequency peak moves for velar stops in varying contexts, let us look again at the trajectories of F2 and F3.

3.3. Trajectory extrapolation

Just as at onset, in the context of the back vowel /a/, the formant frequencies for the voiced stops are tightly grouped, and can be easily classified; in the front vowel /i/ context, though, the clusters are closer which implies a marginally higher probability of classification error. Nevertheless, the average release values were calculated for the F1, F2 and F3 trajectories by extrapolation, and are superimposed on the burst spectra in Figure 3. The formant measurements were less reliable for F4 and so a direct mean was taken, rather than using the trajectory models. In the case of the velar bursts, therefore, F3 is very low and close to F2 for the back vowel, whereas both are high for the front vowel, accounting for the shift we saw earlier.

4. Discussion

In calculating the ensemble spectra, we tried using the periodic and aperiodic components, in turn, but found that artefacts introduced at voice onset gave misleading results, so for this part the original recording was employed, in preference. Having centred the frames on the burst epoch, the averaged magnitude spectra were essentially identical to the averaged power spectra, in all three cases. The PSHF performs best during sustained phonation and, while it has been shown to be robust to typical jitter and shimmer fluctuations [3, 5], it has difficulty at voice onset where the rapid changes affect the estimated components for a couple of pitch periods either side. In these syllable-initial voiced plosives, for which voicing starts at the same time as or shortly after release, the PSHF is thus handicapped, and its decomposition does not help analysis in these regions. However, were voicing to continue through closure of the stop consonant, as can occur in syllable-medial context, we might expect this problem to be diminished, and indeed would like to see this demonstrated in future experiments. Nonetheless, the PSHF was helpful in cleaning up the component attributable to voicing with respect to the original recording, so as to improve the formant analysis using the periodic component.

Optimising the fit between the formant measurements and the trajectory function by the method of least squares, our results demonstrated that the exponential trajectory outperformed polynomial ones of equal or fewer degrees of freedom, much as anticipated. In other contexts, such as /ugo/, where the trajectory does not resemble an exponential function [13], a model with more than one pole may suffice. Formant estimates at release had similar discriminatory properties to those at onset as strong place cues, but they are both highly dependent on vowel context. Although the present trajectories on their own are far from definitive for purposes of classification, the similarities across voiced/voiceless pairs suggest parameters may be tied on the basis of place for more-robust training.

Similarly, the plosives' spectral charateristics were related by place, and not significantly affected by presence or absence of voicing. Minor differences occurred as a result of vowel context, except for velar plosives, where [ki,gi] were substantially different from [ka,ga]. Here, the formants predicted at release corresponded to high-amplitude regions of the burst spectra.

In conclusion, having used the PSHF to provide a cleaner voicing signal, exponential functions were fitted to measured formant-frequency tracks at the beginning of vowels that followed the plosive consonants. These functions were traced back to the time of plosive release to predict the formant frequencies during the excitation. Ensembles of plosives burst spectra were averaged in order to identify their pertinent features, with particular attention to the spectral peaks and troughs (corresponding to poles and zeros in the vocal-tract transfer function). The formant values at release appeared to provide a link between the characteristics of the burst spectrum and their ensuing development. Thus, having segregated these place cues on the basis of their acoustic "texture" (i.e., one is transient, the other periodic), we were able to extract more representative features for each, which in combination may provide a more robust approach to discrimination.

5. Acknowledgements

The author would like to thank Dr Christine Shadle for helpful discussions, Li Qun for help with the speech recordings, and Prof Martin Russell for his encouragement and support.

6. References

[1] K. N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1998.

[2] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am.*, vol. 64, no. 5, pp. 1358–1368, 1978.

[3] P. J. B. Jackson, *Characterisation of plosive, fricative and aspiration components in speech production*, Ph.D. thesis, Dept. Electronics & Comp. Sci., Univ. of Southampton, UK, 2000, http://web.bham.ac.uk/p.jackson/abstracts.html.

[4] P. J. B. Jackson and C. H. Shadle, "Frication noise modulated by voicing, as revealed by pitch-scaled decomposition," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1421–1434, 2000.

[5] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Trans. Spch. Aud. Proc.*, due to appear 2001.

[6] P.C. Delattre, A.M. Liberman, and F.S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Am.*, vol. 25, no. 4, pp. 769–773, 1955.

[7] W. J. Holmes and M. J. Russell, "Probablistic-trajectory segmental HMMs," *Comp. Speech & Lang.*, vol. 13, no. 1, pp. 3–37, 1999.

[8] A. Wiewiorka and D. M. Brookes, "Exponential interpolation of states in a hidden Markov model," *Proc. Inst. of Acoust.*, vol. 18, no. 9, pp. 201–208, 1996.

[9] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for vocal-tract-resonance dynamics," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[10] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987.

[11] H. M. Sussman, D. Fruchter, and A. Cable, "Locus equations derived from compensatory articulation," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3112–3124, 1995.

[12] M. A. Huckvale, *Speech Filing System*, SFS Win v1.22, Dept. Phon & Ling. Univ. Coll. London, UK, 2000, http://www.phon.ucl.ac.uk/resource/sfs/.

[13] D. Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," *J. Acoust. Soc. Am.*, vol. 72, no. 2, pp. 379–389, 1982.