

Department of Electronic Engineering

**Automated Vowel Detection & Recognition System**

Daniel Jackson

ELE0074M Voice: Acoustics & Applications

January 2021

1. Introduction & Background

In phonetics, vowels are phonemes made from glottal vibrations with an open vocal tract [1]. The type of vowel is determined by the shape of this vocal tract, which is affected by the positioning of the lips, jaw and tongue [2]. These articulators adjust the resonant frequencies of the vocal tract, known as formants, which differentiate vowels from one another [3]. In continuous speech, the start of a vowel is known as the Vowel Onset Point (VOP). This report will detail the design of a software system for the automatic detection of vowels within continuous English speech and the recognition of that vowel. In English, there are two types of vowels: monophthongs and diphthongs. Monophthongs are vowels which remain constant throughout, whereas diphthongs change from one vowel to another during their production [3]. The system detailed in this report will focus only on the analysis of monophthongs.

There is a large amount of research surrounding VOP detection, and many different methods used for this. One popular method is to extract the excitation source using Linear Predictive Coding (LPC) coefficients. LPC is used to predict the next point in a signal as a linear combination of the previous values, and the LPC coefficients are used to tune this equation [4]. Equation 1 shows a representation for LPC, where is the next sample, is the order of the predictor, are the LPC coefficients and is the residual signal, which acts as the excitation source.

The coefficients are calculated by minimising the energy of the residual signal [4]. They are calculated for short frames of audio and are then used to generate an inverse filter which the audio is processed with. This filtering leaves the Linear Predictive (LP) residual, which contains the excitation source information of the vowel [2]. The Hilbert Envelope of that residual is then calculated, and the first-order differential of the envelope is then calculated using the First-Order Difference (FOD) operation [2, 5]. Finally, this differential is convolved with the First Order Gaussian Difference (FOGD) operator, which is the equivalent to applying a Gaussian window in the frequency domain [2, 5]. This effectively leaves a smoothed, accentuated, time-domain representation of the excitation source. The peaks from this signal can then be picked to determine where the VOPs occur within the audio. Figure 1 shows an example of this, with the original waveform and labelled VOPs compared to the representation of the excitation source and the automatically detected VOPs.

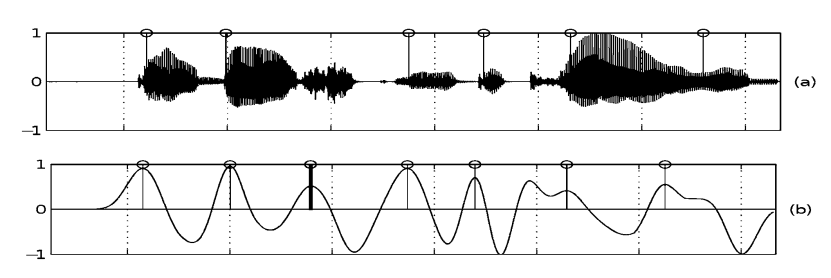


Figure 1: Example of VOP detection using LPC coefficients to extract the excitation source information. (a) Shows the original speech signal with the VOPs labelled, and (b) shows the extracted excitation source signal with automatically detected VOPs using peak picking [5].

An alternative method for detecting VOP is to extract the spectral peaks energy over an audio signal, and then pick the peaks from this signal. This can be done by performing a Short Time Fourier Transfer (STFT) over the length of a signal, and for each frame of summing the amplitudes of the 10 largest peaks in the frequency domain [5]. As with the LPC approach, this can then be smoothed and accentuating by calculating the first-order differential using the FOD operation and then convolving with the FOGD operator [5]. This signal can then be peak picked to find the VOPs. Figure 2 shows the spectral peaks signal along with the original audio signal.



Figure 2: Example plots of original audio signal (top) and the spectral peaks energy signal (bottom).

There are also different methods for automatically identifying formants within vowels. One of the most popular methods for doing this also makes use of LPC. This method is done by calculating the LPC coefficients for a short frame of audio, and then calculating the filter frequency response curve of these coefficients [6]. This effectively gives an approximation of the spectral envelope of the signal, sometimes called the LPC spectrum [7]. The formant frequency can then be estimated by picking the peaks of this signal [6]. An example of this is shown in Figure 3.



Figure 3: Plot of the LPC spectrum against the original spectrum with the formants peak picked. The first 2 formant frequencies are 667 Hz and 1830 Hz respectively, and the vowel being measured is /æ/ spoken with a male voice.

Alternatively, formant frequencies can simply be peak picked from the frequency spectrum following an FFT transform. However, it can be challenging to pick the correct peak due to the many peaks usually present, making it easy to get the incorrect formant frequencies in an automated system.

Based on the first two formant frequencies detected within a vowel, it is possible to estimate what vowel it is from measured ranges. One of the most used pieces of research into the formant frequencies within vowels is Peterson and Barney’s study from 1952 [8]. Figure 4 shows a plot of the first formant () against the second formant (), with the measured ranges of certain English vowels labelled.

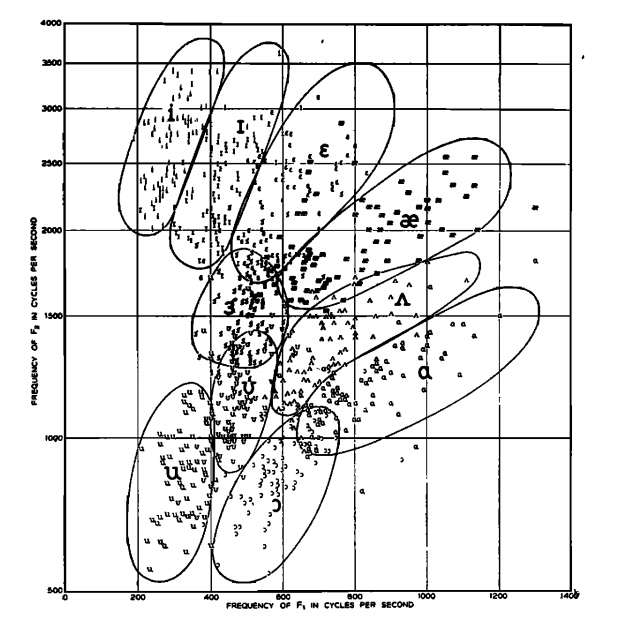


Figure 4: Frequency of first formant against second formant for 10 English vowels [8].

1. System Design

The system detailed in this report aims to combine some of the approaches mentioned in the previous section to build the full system. The system is implemented in MATLAB and makes use of the Signal Processing Toolbox. All the code is included in the appendix, as well as in the supporting files.

Before VOP detection or vowel estimation is attempted, there is a small amount of pre-processing applied to the signal. Firstly, if the signal is a stereo recording, it is converted to mono by summing the two channels into a single channel. The signal is then normalized between 1 and -1, and a 5ms fade-in and fade-out is added.

The vowel onset detection of the system is implemented using the spectral peaks energy approach described in the previous section. It is worth noting that for this system the focus was on capturing a part of the vowel for formant analysis, and so the success criteria is just for the detected VOP to fall within the duration of the vowel, as opposed to falling directly at the vowel’s onset. The signal is split into frames which are 2048 samples long, and each of these frames is windowed using a hamming window of the same size. The windowing is used to reduce spectral leakage. There is a 1024 sample overlap between each frame to avoid losing information due to the windowing. The frequency response of the windowed frame is then calculated by taking the magnitude of a Discrete Fourier Transform (DFT), implemented as a Fast Fourier Transform (FFT). The largest 10 peaks are then picked from this frequency response and summed for each frame. The first-order differential is then calculated, and an 8-point Gaussian window is applied in the frequency domain. Finally, all values below 0 are clipped to 0, as VOP detection only requires the positive peaks.

Once the spectral peaks energy of the signal is calculated, the VOPs are detected. First, the peaks are picked from this signal, with a minimum distance of approximately 200ms between the peaks. This minimum distance was chosen because the average vowel length in American English has been measured as 160ms and 168ms for male and female voices respectively, and the system focuses exclusively on monophthongs, and simply wishes to capture part of the vowel [9]. Therefore, 200ms should offer more than enough range for capturing all vowels, whilst avoiding the same vowel being detected multiple times from neighbouring peaks. Once the peaks have been picked from the signal, the frequency response is carried out at each peak location. Frequencies below 150 Hz are removed from the frequency response, and the maximum value in the spectrum is found. If the maximum value in the spectrum is above 1500 Hz, then the peak is discarded, and it is assumed that this is not a vowel. This is because the F1 should have the highest amplitude and should fall between 150 and 1500 Hz [8]. Doing this helps to reduce false positives during VOP detection.

Once the VOPs are detected, the sample points at which these occur are then used to estimate the formants. At each VOP, a 2048 sample frame of the audio signal is taken, and the LPC coefficients are calculated for this frame, using a 50th-order LP model. The filter frequency response curve is then calculated for these coefficients. This frequency response is then peak picked to find the first three formants of the vowel, as was shown in Figure 3.

Once the vowel formants have been detected, the first two can be used to estimate the vowel. A set of average vowel positions are stored within a matrix. These averages are calculated as the mean between the male and female average measurements in the Peterson & Barney study. The values are shown in Table 1. The distance between the formant positions of the vowel and the average vowel positions are calculated using the Pythagorean Theorem, and the one with the shortest distance is assumed to be the current vowel.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | /i/ | /ɪ/ | /ɛ/ | /æ/ | /ʌ / | /ɑ/ | /ɔ/ | /u/ | /ʊ/ |
|  | 290 | 410 | 570 | 760 | 760 | 790 | 580 | 335 | 455 |
|  | 2540 | 2235 | 2085 | 1885 | 1295 | 1155 | 880 | 910 | 1090 |

Table 1: Average vowel values used within the system for estimating vowels based on their first two formants.

1. Results & System Success

The success of the system will be analysed separately for each section. For testing and gathering the results, five recordings of the phrase “Please call Stella” in different accents will be used. These recordings are sourced from The Speech Accent Archive and are included in the supporting files of this report, named according to the accent of the speech [10].

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Accent* | *Vowel 1* | | *Vowel 2* | | *Vowel 3* | | *Vowel 4* | |
| *Start* | *End* | *Start* | *End* | *Start* | *End* | *Start* | *End* |
| Dublin | 0.196s | 0.300s | 0.463s | 0.591s | 0.750s | 0.814s | 0.875s | 0.942s |
| London | 0.074s | 0.153s | 0.297s | 0.400s | 0.641s | 0.713s | 0.759s | 0.904s |
| Brooklyn 1 | 0.165s | 0.251s | 0.409s | 0.510s | 0.700s | 0.783s | 0.866s | 1.003s |
| Brooklyn 2 | 0.178s | 0.313s | 0.499s | 0.648s | 0.875s | 0.996s | 1.118s | 1.229s |
| Glasgow | 0.216s | 0.305s | 0.468s | 0.574s | 0.793s | 0.896s | 0.945s | 1.099s |

Table 2: Manually measured vowel onset and offset times in the five audio clips used for testing.

The criterion for success of the VOP detection is that the detection lies within the duration of the vowel. The criteria for failures of the VOP detection are missing a VOP and detecting a VOP where there is no vowel. The vowel onsets and offsets were then manually measured from the spectrograms of the respective waveforms using PRAAT. The onset and offset points for the four vowels in each clip are shown in Table 2.

The 20 VOPs automatically detected by the system for the same clips are shown in Table 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Accent* | *VOP 1* | *VOP 2* | *VOP 3* | *VOP 4* |
| Dublin | 0.279s | 0.557s | 0.859s | MISSED |
| London | MISSED | 0.372s | 0.673s | 0.859s |
| Brooklyn 1 | 0.232s | 0.488s | 0.766s | MISSED |
| Brooklyn 2 | 0.232s | 0.581s | 0.952s | MISSED |
| Glasgow | 0.232s | 0.557s | 0.882s | MISSED |

Table 3: VOPs detected automatically by the system from five audio clips used for testing. Green cells represent correctly identified VOPs, and red cells represent missed VOPs.

Of these 20 vowel positions, 14 were correctly identified, 6 were missed, and there were 2 false detections. Therefore, the accuracy of the system is 70%, and the false positive rate is 10%.

Now the accuracy of the formant detection will be analysed. Only the correctly identified VOPs will be used for this analysis. The first two vowel formants were measured using PRAAT, and these values are shown in Table 4.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Accent* | *Vowel 1* | | *Vowel 2* | | *Vowel 3* | | *Vowel 4* | |
| *F1 (Hz)* | *F2 (Hz)* | *F1 (Hz)* | *F2 (Hz)* | *F1 (Hz)* | *F2 (Hz)* | *F1 (Hz)* | *F2 (Hz)* |
| Dublin | 280 | 2420 | 555 | 1100 | - | - | - | - |
| London | - | - | 385 | 885 | 635 | 1780 | 410 | 1475 |
| Brooklyn 1 | 320 | 2120 | 550 | 940 | 545 | 1530 | - | - |
| Brooklyn 2 | 430 | 2660 | 540 | 1160 | 660 | 1830 | - | - |
| Glasgow | 280 | 2420 | 510 | 980 | 600 | 1830 | - | - |

Table 4: Measured vowel formants from the five audio clips used for testing. Vowels which were not detected by the system have been left blank here.

The formant values automatically detected by the system for the same clips are shown in Table 5.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Accent* | *Vowel 1* | | *Vowel 2* | | *Vowel 3* | | *Vowel 4* | |
| *F1 (Hz)* | *F2 (Hz)* | *F1 (Hz)* | *F2 (Hz)* | *F1 (Hz)* | *F2 (Hz)* | *F1 (Hz)* | *F2 (Hz)* |
| Dublin | 242 | 2401 | 264 | 1024 | - | - | - | - |
| London | - | - | 237 | 2525 | 1674 | 2584 | 86 | 3720 |
| Brooklyn 1 | 226 | 2083 | 345 | 877 | 511 | 2465 | - | - |
| Brooklyn 2 | 334 | 2724 | 517 | 1050 | 673 | 1545 | - | - |
| Glasgow | 232 | 2401 | 355 | 670 | 592 | 1459 | - | - |

Table 5: Vowel formants automatically detected by the system for the five audio clips used for testing.

Comparing results from both tables shows the difference between the manually measured and automatically extracted formants. There is a mean difference of 390.1 Hz between the measured and extracted formant, and a lot of variation between the vowels. The smallest difference between the measured and extracted formants is 13 Hz, and the greatest difference is 1640 Hz, and the standard deviation of the differences is 578.6 Hz. This implies that the formant extraction within the system is not very accurate. This is likely due to the wrong peaks being picked from the LPC spectrum. Some evidence of this can be seen in Figure 6, where it is likely happening due to their being very poorly defined peaks in the spectrum.



Figure 6: Formants incorrectly extracted from peak picking due to poorly defined peaks in the LPC spectrum. This LPC spectrum is from the 4th vowel in the testing audio clip for the London accent.

Finally, the vowel’s estimated by the system will be compared with the actual spoken vowels. The actual spoken vowel will be determined by using the manually measured formants and reading the labelled vowel from the chart shown in Figure 4. The actual spoken vowels are shown in Table 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Accent* | *Vowel 1* | *Vowel 2* | *Vowel 3* | *Vowel 4* |
| Dublin | /i/ | /ʊ/ | - | - |
| London | - | /u/ | /æ/ | /ʊ/ |
| Brooklyn 1 | /i/ | /ʊ/ | /ɛ/ | - |
| Brooklyn 2 | /ɪ/ | /ʊ/ | /æ/ | - |
| Scotland | /i/ | /ʊ/ | /ɛ/ | - |

Table 6: Actual measured vowels from the formants of the vowels within each of the test audio clips.

The estimated vowels are shown in Table 7.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Accent* | *Vowel 1* | *Vowel 2* | *Vowel 3* | *Vowel 4* |
| Dublin | /i/ | /u/ | - | - |
| London | - | /i/ | /æ/ | /i/ |
| Brooklyn 1 | /ɪ/ | /u/ | /i/ | - |
| Brooklyn 2 | /i/ | /ʊ/ | /ʌ/ | - |
| Scotland | /i/ | /u/ | /ɛ/ | - |

Table 7: Vowels automatically detected by the system for each of the five test audio clips. Cells shaded green estimated the vowel correctly, and cells shaded red estimated the vowel incorrectly.

This shows that only 5 out of the total 14 vowels were estimated correctly, giving an accuracy of just 35.7%. This is largely due to the inaccuracies in the formant extraction, but also relates to the close similarities between some vowels and the method of detection. For example, considering the 3rd vowel from the Brooklyn (b) example, the detected formants fall in a region of overlap on the Peterson and Barney chart between /ʌ/ and /æ/, but are closer to the centre point of /ʌ/. Therefore, the chosen vowel is /ʌ/, despite the formants also matching the formant frequencies of the correct vowel.

1. Conclusion & Further Work

Different sections of the system have been shown to have varying levels of success through their analysis. The Vowel Onset detection has shown reasonably good success, although this could be fine-tuned for better levels of accuracy. However, enough accuracy has been shown to demonstrate that the spectral peaks energy method of vowel onset detection is effective, reflecting what was shown in the literature [2, 5]. The formant extraction of the system performed poorly, as mentioned in the previous section, with a large average difference between the formants. This is likely due to poorly defined peaks in the LPC spectrum when the formant peaks are picked. Improving this accuracy could be approached in a couple of ways: the order of the LP model could be increased, to improve the definition of the spectrum’s peaks; and multiple windows could be taken around the VOP and the detected formants at each of these windows averaged out. Finally, the vowel estimation of the system also performed poorly, achieving an accuracy of only 35.7%. This is mostly due to the inaccuracies in the formant extraction, as where the formants were extracted most accurately, the vowel estimation was correct. Overall, I believe that improving the formant extraction of the system would greatly improve its overall success at estimating vowels. It is also worth noting that the results and analysis were carried out using a very small set of test data, and further testing with a larger data set would be necessary for definitive conclusions to be drawn.

1. References
2. A. Cruttenden, *Gimson’s Pronunciation of English, 7th Edition*. Hodder Education, 2008.
3. S. R. Mahadeva Prasanna et al., ‘Vowel Onset Point Detection Using Source, Spectral Peaks, and Modulation Spectrum Energies’, *IEEE Trans. Audio Speech Lang. Processing*, vol. 17, no. 4, pp. 556–565, May 2009.
4. Plural Publishing, Incorporated and D. Murphy, *Voice Science, Acoustics, and Recording*. San Diego, UNITED STATES: Plural Publishing, Incorporated, 2007.
5. D. Ellis. (2013). *Lecture 6: Linear Prediction*. [Online]. Available: https://www.ee.columbia.edu/~dpwe/e4896/lectures/E4896-L06.pdf. [Accessed: 06 Jan. 2021].
6. A. K. Vuppala et al., ‘Vowel Onset Point Detection for Low Bit Rate Coded Speech’, IEEE Trans. Audio Speech Lang. Processing, vol. 20, no. 6, pp. 1894–1903, Aug. 2012.
7. R. C. Snell and F. Milinazzo, ‘Formant location from LPC analysis data’, *IEEE Trans. Audio Speech Lang. Processing*, vol. 1, no. 2, pp. 129–134, Apr. 1993.
8. J. O. Smith, ‘Spectral Envelope Linear Predictors’, in Spectral Audio Signal Processing, 2011 [Online]. Available: <https://ccrma.stanford.edu/~jos/sasp/Spectral_Envelope_Linear_Prediction.html>.
9. G. E. Peterson and H. L. Barney, ‘Control Methods Used in a Study of the Vowels’, Journal of the Acoustical Society of America, vol. 24, no. 2, pp. 175–184, Mar. 1952.
10. E. Jacewicz et al., ‘Vowel Duration in Three American English Dialects’, *Am. Speech*, vol. 82, no. 4, pp. 367–385, 2007.
11. S. Weinberger. (2015). *Speech Accent Archive*. [Online]. Available: http://accent.gmu.edu. [Accessed: 09 Jan. 2021].
12. Appendix



Figure 7: MATLAB code from analysevowels.m.



Figure 8: MATLAB code from monoconvert.m



Figure 9: MATLAB code from dualfade.m.



Figure 10: MATLAB code from spectralpeaks.m.



Figure 11: MATLAB code from vowelpeaks.m.



Figure 12: MATLAB code from vowelformants.m.



Figure 13: MATLAB code from estimatevowels.m