VOWEL LANDMARK DETECTION
Andrew Wilson Howitt
Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 USA
(617)253-5957 voice (617)258-7864 fax
howitt@mit.edu http://www.mit.edu/~howitt/

Abstract

Landmark based speech processing is a component of Lexical Access From Features (LAFF), a novel paradigm for feature based speech recognition. Detection and classification of landmarks is a crucial first step in a LAFF system. This work tests the theoretical characteristics of vowels, and shows results for work in progress on a Vowel Landmark Detector.

Acoustic theory predicts first formant peaks in vowels, both in frequency and amplitude (at least for vowels between orally closed consonants). Formant tracking measurements found peaks in about 94% of vowels in the TIMIT database. Vowels which do not show a peak generally do not obey the theoretical assumptions, or are liable to formant tracker error due to nasalization, glottalization, or aspiration. Amplitude peaks are more reliable than frequency peaks. Peaks tend to occur early in the vowel, and frequency peaks tend to occur slightly before amplitude peaks. A fixed spectral band gave performance comparable to the formant tracker for this task, allowing a simpler detection algorithm.

Previous work on a Vowel Landmark Detector is extended by use of a multilayer perceptron (MLP) to combine knowledge-based acoustic cues. The MLP decreases error rate to about 12%, of which about 8% are deletions. Since about 6% of vowels had no detectable peak, this performance is close to the expected limit of a peak picking algorithm. Work is continuing on algorithm improvements, including the output of confidence scores.

1 Introduction

The primary motivation for this work is to detect Vowel landmarks as part of the front end of a LAFF speech recognition system. LAFF [6, 7] is a knowledge-based approach to speech recognition, in which landmarks (indicating vowels, consonants, or glides) are detected in the speech signal, and phonetic features are detected and attached to the landmarks. Landmark detectors for glides and consonants [3] have already been developed, leaving only vowel landmarks yet to be done.

According to acoustic perturbation theory [8, p. 148], the first formant should be reduced in frequency whenever a constriction is made in the front half of the vocal tract. Since all English consonants are articulated in the front half of the vocal tract, we expect that all consonants will cause F1 to drop in frequency (in English). Conversely, F1 frequency is maximally high when the front half of the vocal tract is maximally open, making the F1 maximum a natural choice for the location of the Vowel landmark. 1

Also according to basic acoustic theory, an increase in the first formant frequency should be accompanied by an increase in amplitude, both of the first formant itself and the overall spectrum [8, p. 154], assuming that the glottal source is fixed in amplitude and spectral shape. To the degree that this holds true, either F1's frequency or amplitude may be used to determine the Vowel landmark.

Two assumptions underlie these predictions. First, the glottal source should be fixed, without major changes in either amplitude or spectral content. Second, the influence of secondary articulation (in particular, nasalization and glottalization) should be minimal. A major change in the glottal source, or major influence of secondary articulators (such as nasalization), will invalidate the predictions.

There are also several pragmatic concerns when performing experiments. For instance, the database labeling may describe an underlying vowel, preceded and followed by underlying consonants, but may not correspond to the acoustic realization of the speech signal (for example, a nasal murmur instead of a vowel, or an elided consonant). Also, errors in the formant tracking algorithm will cause incorrect values for F1 frequency, which may also affect the computation of F1 amplitude.

2 Experimental Methodology

A series of experimental studies was performed to test the theoretical predictions. The TIMIT database was used for these experiments. The entire database was used (all utterances from all talkers and dialects, both training and test sets) for a total of 80856 vowels by 630 talkers (10 utterances each).

Formant frequency tracks were extracted using the Entropic ESPS formant program, spaced every 5 ms, with LPC resonances to find formant frequencies, and dynamic time warping to group the formant tracks. Formant amplitude tracks were generated from a spectrogram using a 6 ms window, 1 ms frame period, smoothed in time over 15 frames, and the amplitude averaged over 100 Hz around

1A possible exception occurs when semivowels like /r/ are adjacent to high vowels like /i/. In these cases, the source is liable to be reduced in the semivowel relative to the vowel, so that F1 amplitude may peak in the vowel even if F1 frequency does not.
Acoustic theory predicts that F1 should show a proper peak in vowels between orally closed consonants (both frequency and amplitude). The assumptions that underlie this prediction are (1) the vowel is orally open, and the neighbors are orally closed, (2) F1 is clear and measurable, without nasalization, glottalization or aspiration, (3) the acoustics are stable, without lateralization, or glottal source change (for amplitude peaks).

The peaks were classified as occurring at beginning, middle, or end of the labeled vowel. The peaks were also classified by context, with the preceding and following segments either orally Closed (stops, fricatives, nasals), orally Open (vowels, semi-vowels) or X (unknown, including flaps, glottals, /l/, and pauses). Amplitude and frequency peaks were studied separately. The theoretical prediction is that the amplitude peak should not appear at the beginning of the vowel in CV- contexts, and should not appear at the end of the vowel in V-C contexts. If the amplitude peak does appear in these unpredicted contexts, we expect that one or more of the underlying assumptions is not true.

Results for amplitude peaks show that, in general, the theoretical prediction holds true. Most vowels have a peak somewhere in the middle (94.2%), and of those which do not, none have the peak at the end (3.73%) than at the beginning (2.03%). Almost no peaks occur in violation of theory (184 or 0.22%).

An automatic procedure was developed to seek evidence of assumption violations for the peaks which do not obey the theoretical prediction. Evidence for assumption violations included adjacent nasal or glottal segments, liquids /r/ or /l/ adjacent to high vowels, and vowels which are extremely short in duration (less than 30 ms). A total of 370 assumption violations were found for the 814 tokens which disobeyed the theoretical prediction, averaging about 2 violations per token. Manual inspection of borderline cases indicated that all had assumption violations that were not found by the automatic procedure (devoicing or aspiration of vowels, and elision of consonants). Only one token had no automatic label at all, and manual inspection showed it to be nasalized and devoiced (the capitalized vowel in “shorten this skirt”).

Results for frequency peaks show that, in general, the theoretical prediction holds true, although not as well as for amplitude peaks. Many vowels have a peak in middle (80.1%), and of those which do not, more have the peak at the end (8.44%) than at the beginning (2.46%). Few peaks occur in violation of theory (1709 or 2.11%). The poorer performance for frequency peaks might be explained by errors in the formant tracker, which might be alleviated by smoothing operations in the amplitude computation.

Male and female talkers scored about the same, indicating that high F0 values were not causing problems for the formant tracker. Some evidence was found for formant tracker failure at vowel boundaries, where spurious values may cause false peaks – frequency values for peaks at the endpoints are significantly higher than for peaks in the middle, and less well correlated with vowel height.

Overall, the experimental results agree with the theoretical prediction. The vast majority of vowel tokens show peaks (in both frequency and amplitude) somewhere in the middle of the vowel. Amplitude peaks are more consistent than frequency peaks.

2.2 Frequency - amplitude coincidence

Acoustic theory predicts that the amplitude and frequency peaks should occur at the same place in time. The assumptions that underlie this prediction are the same as in the previous experiment, plus stability in the voicing source (instability would affect the amplitude peak).

For this experiment, only vowels with both amplitude and frequency peaks in the middle were used (690 tokens out of 808.56). For each vowel, the peak location was computed as a percentage of the vowel’s duration. In this way, the peak locations were normalized for duration, and histograms of peak location in normalized duration were plotted (figures 1 and 2). These figures show that peaks
Both frequency and amplitude peaks tend to appear early in the vowel rather than late, and that the frequency peaks tend to appear earlier than the amplitude peaks.

To examine the data further, a histogram of tokens' separation from the diagonal was computed. We expect the resulting “difference diagonal” histogram to have a high peak at zero, and taper off rapidly on both sides. The result, shown in figure 3, validates the basic prediction, as the data show a high peak on the diagonal, and taper off rapidly for tokens off the diagonal. However, the figure also shows a pronounced “shoulder” just below zero on the horizontal axis, indicating a tendency for frequency peaks to occur earlier than amplitude peaks.

Although most of the tokens are close to the diagonal in figure 3, there are a significant number off the diagonal, in violation of the theoretical prediction. It may be surmised that these anomalous tokens should show violations of the assumptions underlying the theoretical prediction, and that there should be more assumption violations (or more extreme violations) for tokens very far from the diagonal. If a histogram of assumption violations per token were plotted in the same way as figure 3, it should show a minimum at zero, and rising values towards both positive and negative extremes.

Violations of theoretical assumptions were detected using the same procedure as in the previous section. The result is shown in figure 4. Again, the basic prediction is validated, as the data show rising values towards both positive and negative extremes.

Overall, the experimental results agree with theory. Both peaks tend to occur early in the vowel, and frequency peaks tend to occur earlier than amplitude peaks. The diagonal histogram shows peaks are highly correlated, and assumption violations increase with distance from the diagonal.

### 2.3 Fixed energy band

The prior two experiments show that amplitude peaks are probably better than frequency peaks for finding vowels. It is desirable to find a fixed energy band for finding amplitude peaks, avoiding the use of a formant tracker. This experiment will see if a fixed energy band can perform as well as formant tracking.

The first experiment (amplitude only) was repeated, using the energy in a band around F1 (300 to 900 Hz). Results were comparable to those derived from use of a formant tracker. Most vowels have a peak somewhere in the middle (96.2%), and of those which do not, more have the peak at the end (2.6%) than at the beginning (1.2%). Almost no peaks occur in violation of theory (257, or 0.32%).

The edges of the frequency band, and the transition widths, were varied to investigate the performance change (where performance is measured both by maximizing vowels found with peaks in the middle, and by minimizing the theoretical prediction violations). The performance was not found to be sensitive to changes in these values, only fractions of a percent change were observed. The insensitivity of performance to changes in the frequency band is a desirable characteristic, allowing robust operation.

### 2.4 Results

The basic results are very encouraging for a Vowel Landmark Detector. In general, the predictions of acoustic theory are supported. Most vowels show an amplitude peak, and those that do not are explainable by conditions such as nasalization, glottalization, and aspiration of vowels, and elision of consonants. A fixed frequency band works well for amplitude peak detection, so no formant tracker is necessary.

### 3 Implementation

In previous work [2], a basic Vowel Landmark Detector (VLD) was developed, based on an algorithm for automatic syllable detection [5]. That VLD used peak picking on the F1 band (300 to 900 Hz) and made decisions based on three cues: peak level, peak-to-dip value, and dip-to-dip duration. It achieved fairly good error rates (13.6%
train, 14.8% test) on the TIMIT database. Most of the errors were deletions (11.6% train, 12.9% test). The results shown in the previous section indicate that about 6% of vowels do not have a peak, and therefore will be deleted by any peak picking technique. The VLD’s deletion rate of about 12% means that there is still room for improvement.

Multiple acoustic cues must be combined for a detection decision. The VLD in [2] used simple thresholds for each of the three cues. More sophisticated techniques for combination of multiple acoustic cues include fuzzy logic [1] and neural networks, such as a Multi-Layer Perceptron (MLP) [4]. For the VLD, an MLP of one hidden layer was chosen. Two hidden units were found to be adequate to capture the variability in the data.

Training of MLPs is usually done with back propagation, which is fast but requires a differentiable error metric. Error rate is the desired quantity to minimize, but is not differentiable. The solution was to use a two stage training process. First, back propagation was used to train the MLP on a sum-of-squares error metric. Then, training was continued using gradient descent to minimize error rate. Simulated annealing was used during gradient descent, to avoid local minima.

The MLP improved the VLD’s error rate (11.8% train, 12.7% test). Again, most of the errors were deletions (8.5% train, 9.23% test). This is an improvement over the previous deletion rate of about 12%, and substantially closer to the 6% predicted by the experiments. Work is ongoing to improve the VLD performance further, and to generate confidence scores to represent the probability of a correct decision.

An example of the VLD’s operation on TIMIT data is shown in figure 5. The low frequency energy track, with landmarks, is shown below the spectrogram. This is a particularly challenging example, showing skewed detection in vowel-semitone-vowel sequences, where the landmark appears in the semitone instead of the vowel. (Scoring techniques for these cases are discussed in [2].) Skewed detections occur at 0.64 s (“you”) and at 1.04 s (“we”), and two-sided skewing occurs at 1.28 s and 1.42 s (first two vowels in “were away”).

4 Conclusion

The basic results are very encouraging for a Vowel Landmark Detector. In general, the predictions of acoustic theory are supported. Most vowels show an amplitude peak, and those that do not are explainable by conditions such as nasalization, glottalization, and aspiration of vowels, and elision of consonants. A fixed frequency band works well for an amplitude peak detection, so no formant tracker is necessary. Performance of the basic detector is improved by use of an MLP to combine the acoustic cues, and approaches the limit derived from the scientific study.

References