Compensating for Vowel Coarticulation in Continuous Speech Recognition

James L. Hieronymus *

Abstract

Coarticulation alters the vowel formant characteristics in continuous speech. Studies of isolated monosyllables in the literature suggest that some phonemes cause more severe distortions than others. The largest changes are caused by /r/, /l/, /w/. Unstressed vowels are most affected. Previous studies by Holmes [6] and by us [13] indicate that these effects are even larger for continuous speech. Vowel recognition algorithms which do not take context into account in continuous speech normally achieve correct recognition of approximately 75% for the three top choices from the recognizer. By developing methods which explicitly model the phonetic context, higher levels of performance are expected to be achieved. An ongoing study is being made of all 16 of the American English vowels using a subset the DARPA acoustic-phonetic data base, a phonetically labeled 6300 sentence data base with 630 talkers. A subset of 7 vowels /iy/, /l/, /eh/, /ae/, /o/ and /u/ have been studied in all major contexts. The formant temporal patterns are being examined for phoneme triples and quintuples with a vowel in the center. These formant patterns are discussed along with some effects of stress and speaking rate.

Introduction

Early studies of vowels by Potter and Steinberg [1] and Peterson and Barney [2] used isolated monosyllabic words to provide canonical formant frequencies for each monothong vowel in American English. Given a neutral h.d context, the vowels clustered well in the space of the first and second formant, across hundreds of talkers. Vowels were believed to be characterized by the steady state frequencies of the first two or three formants.

This view was challenged by some perceptual studies by Fairbanks and Grubb [3] for American English and Fujimura and Ochiai [4] for Japanese. They showed that the steady state portions of the vowels excised from the words were less identifiable than the whole words. In some way vowel context is very important in disambiguating vowels. Strange et al [5] showed that by listening to just the transitions into and out of the vowels with silence replacing the usual vowel, listeners were able to identify the vowel almost as well as when played the whole word. An important cue to identification in this experiment was duration. With duration not available as a cue, the vowel-less vowel identification accuracy fell to somewhere between isolated vowels and whole word.

An early study by Shearme and Holmes [6] showed that vowels in continuous speech very seldom had steady states and often did not overlap the Peterson-Barney contours in any part of their frequency trajectories in time. Generally the vowels are much more centralized in continuous speech and the vowel formant regions overlap considerably due to coarticulation.

The durations of vowels are longer for isolated words than for continuous speech. Vowels in front of pauses are also lengthened. One possible hypothesis from this data is that vowels in continuous speech are more identifiable than vowels in words bounded by silence. Presumably this is because a great deal of information is carried in the formant transitions.

Strange et al [5] have advanced the theory that vowels are completely determined by the formant transitions into and out of the vowel. Nearey and Assmann [7] have postulated three alternate possible explanations for vowel perception, dual formant target, target plus slope and target plus direction. In their experiments they were unable to determine which mechanism was a clear best choice. Experiments of Gay [8] would seem to favor target plus slope, those of Pols [9] favor target plus direction for Dutch diphthongs. All of these studies were conducted on isolated words.

The approach presented here is to attempt speaker independent vowel recognition based on formant frequencies and pitch. Speaker normalization using formant differences and F1 - pitch as suggested by Syrdal and Gopal [10] has been studied for this data. The second source of variation

* Institute for Computer Sciences and Technology, Building 225 Room A216, U. S. National Bureau of Standards, Gaithersburg, Md. 20899, U.S.A.
is the vowel context. Coarticulation causes very large formant variations. The semi-vowels /r/ /l/ /w/ and /y/ have the strongest influence on nearby vowels. We represent the vowel and its environment by the vowel and its nearest neighbor phonemes (vowel centered triplet). This is taken as the minimum unit of coarticulation, with some notable exceptions. Thus a vowel with a following /r/, for instance, will be severely affected. The exceptions are when a labial consonant is the nearest neighbor to the vowel and the semi-vowel is the next nearest neighbor. Thus the labial stops /b/ and /p/, the labial and labiodental fricatives /v/ and /f/, labial nasal /m/ all transmit coarticulation. Thus, we also are studying vowel centered quintuples.

**Theory**

Our view of vowels is that the composite formant trajectory of the vowel and its neighbors are what allow the human to uniquely identify the vowel. How much of the trajectory is needed? Generally enough to cover the phonemes which cause severe coarticulation. In some cases, vowel centered triples are sufficient. Phillips [11] in a very recent study of human perception of triphones, shows larger units are needed. Perhaps quintuples give enough context. A definitive answer awaits further study.

Thus a representation of the formant trajectory which gives a small number of parameters, yet approximates the formant trajectory well is needed. We represent the three formant trajectories for a vowel centered triplet by dividing the region into thirds in time, and finding the mid value and slope in the center of each third. These parameters should be sufficient to identify the vowel based on triplet context. Senelff [12] has a vowel recognition scheme which fits lines to auditory spectrogram dark regions without selecting formants. There are similarities between her approach and ours.

**Method**

In order to parameterize formant trajectories we first automatically find the formants using anharmonic pitch synchronous Fourier Transforms and a knowledge based formant tracker. This formant tracker has an error rate of approximately 5% in a test on 250 sentences by 112 talkers (74 males and 38 females). Next the formant trajectories are fit to beta splines using a least squared error criterion. The splines are then used to determine a mid value and slope for the first third, middle third and last third (in time) of the vowel centered triplet. Thus eighteen parameters are obtained for each triplet, representing line segment formants. Figure 1 shows the resulting formant spline values superimposed on a pitch synchronous spectrogram, with the thirds mid values shown as points. The modeling of the formant trajectories is remarkably good.

The data for the pilot study consists of a subset of 255 sentences from the DARPA Acoustic-phonetic data base which was designed collectively, collected from talkers reading sentences in a quiet environment by Texas Instruments, and hand labeled by the MIT Group. Ultimately the data base will contain 6300 sentences by 630 talkers of American English. All of the major regional dialects are represented. The vowels /iy/ /ih/ /ae/ /eh/ /ow/ /aa/ /u/ were studied. Since the back vowel /uw/ is relatively rare, a special effort was made to include an adequate number. Vowel centered triplets were formed from the hand label boundaries. This procedure will be replaced by an acoustic segmenter for the recognition experiments.

The resulting formant and formant slope data was used to create two dimensional scatter plots of the data. Outliers in these plots were examined and removed if they were the result of formant tracking errors. Male and female data were examined separately. A simple pitch range
test can separate most males from females and high pitched males. High pitched males seem to have formant ranges closer to those of females. Differences of formants which are recommended for talker normalization were also used for scatter plots. Finally the data was used as input to a k-means clustering algorithm, to determine any natural clustering in the space of the formant value and slope data. This gives an indication of the performance of a vowel recognition algorithm based on these parameters.

Discussion of the Data

The durations of the vowels were computed from the hand label boundaries. Figure 2 shows a histogram of duration for six vowels.

![Duration Histogram of 6 Vowels](image)

As a careful examination shows, there is considerable overlap in duration for the tense-lax vowel pairs. Thus for these sentences, there is no clean difference in vowel duration between these types of vowels. Duration may still be useful in partially separating these pairs.

The vowels were plotted on a scatter plot of F1-pitch and F2-F1 for vowels in all contexts except a neighboring semi-vowel. The results are shown in Figure 3.

![Plot of F2 - F1 v.s. F1 - F0 for Males](image)

The rectangular boxes represent regions containing approximately 93% of the /iy/ /ih/ and /eh/ vowels. The overlap of these regions is very large. We conclude that even for this restricted context, average, maximum or midpoint formant frequencies will not be sufficient for vowel identification.

For males the vowel centered triples were k-means clustered using values of the formant differences for the three time regions, individual formant slopes and duration. This resulted in 19
parameters. The results of this clustering for non-semivowel context are shown in Figure 4 for 16 clusters. There are 70 - /iy/, 54 - /ih/, 50 - /ae/, 43 - /eh/, 41 - /uw/ and 33 - /ow/ vowels in this cluster.

Figure 4: Results of K-mean clustering on male vowels

The bottom nodes still contain considerable mixing of vowel identities. The vowel /iy/ seems to appear in most of the clusters. We have every reason to believe that these mixtures represent real coarticulation and speaker effects. Without the slope data the results were somewhat worse.

We plan to use bark scaling for the frequencies and a speaker adaptive scheme which estimates the speaker's formant range from a training sentence. This should control the speaker variability in formants to a large degree. We will conduct tests of the vowel recognition capability of the k-means clusters using variance weighted Euclidean distance measures in the near future.

Acknowledgement

The programming assistance of Bill Majurski and Mike Garris at NBS is greatly appreciated. Thanks to Bill Fisher and George Doddington of Texas Instruments for an excellent job of data collection, to Victor Zue, Stephanie Seneff, Hong Leung and coworkers for phonetically labeling the data. The Spire and Search speech analysis tools created by David Shipman, Scott Cyphers and Mark Randolph at MIT under Victor Zue were extensively used in this project. Without these tools this work could not have been carried out. Discussions with Victor Zue, Stephanie Seneff, Mike Phillips, Roy Gengel and Bill Majurski were very helpful throughout this work. This work was funded in part by DARPA Contract N0003984PD41304.

References