Automated Measurement of Vowel Formants in the Buckeye Corpus

Yao Yao[1], Sam Tilsen[2], Ronald L. Sprouse[1], and Keith Johnson[1]

[1]*University of California, Berkeley*
[2]*University of Southern California*

Abstract: In recent years, corpus phonetics has become a rapidly expanding field. However, the lack of appropriate tools for automatic acoustic analysis hinders further development of the field. In this paper, we present a methodological study on the automatic extraction of vowel formants using both robust linear predictive coding (RLPC; Lee, 1988) and dynamic formant tracking (Talkin, 1987). Acoustic data were taken from the Buckeye corpus of English conversations. We varied two aspects of the analysis - preemphasis and LPC order - to optimize formant tracking results by speaker and vowel. We also show, based on the optimal results, the distribution of ten English vowels in the F1/F2 space in conversational speech.

Keywords: speech corpus, automatic acoustic analysis, vowel formants, Robust LPC

1. Introduction

1.1. Background

With the development of various speech corpora in recent years (e.g. Switchboard corpus of telephone speech, TIMIT speech database, Buckeye corpus of conversational speech), a new trend has emerged in the field of phonetic research, which involves large-scale quantitative analysis of acoustic corpus data. Compared with experimental methods, this new line of research, which is often termed "corpus phonetics", features the use of larger and more realistic phonetic datasets as well as more sophisticated data analysis. Over the past couple decades, the new methodology has produced a fast growing body of literature on various topics including phonetic variation (Byrd 1994; Keating et al, 1994; Raymond et al, 2006; Bell et al., 2003, 2009, etc), speech tempo (Fosler-Lussier and Morgan, 1999; Yuan et al., 2006; Jacewicz et al., 2009, etc), disfluency in natural speech (Shriberg, 2001, etc) and so on.

A classic paper by Keating et al. (1994) presented two case studies using the TIMIT speech database. One of their studies was based on the non-audio part of the corpus (i.e. speaker information, temporal and segmental transcriptions) and investigated durational variation and vowel alternation in the pronunciation of the word *the.* The other study analyzed audio data from the corpus and tested the effect of following vowels on the articulation of velar stops. However, this balanced use of audio and non-audio data has not been pursued in later studies, as non-audio transcriptions have been far more frequently consulted than audio data. As a result, most corpus analyses have concentrated on segmental duration and alternation, but little has been learned about more fine-grained phonetic detail, such as VOT and vowel formants.

The scarcity of corpus acoustic analysis is tied to the lack of appropriate tools for extracting acoustic information from speech corpora. A major challenge is the variability in the signal. Compared with single word production, words in connected speech have more variable forms due to the influence of context. In addition to the phonetic context, predictability also plays an important role: words in more predictable contexts are more prone to phonetic reduction than those in less predictable contexts (Bell et al, 2003, 2009). Moreover, a speech corpus usually contains speech samples from multiple speakers, which will introduce a great deal of inter-speaker variation, due to physiological differences, social factors and idiosyncratic articulatory patterns in individual speakers. Thus an ideal analysis should work reasonably well across a wide range of conditions. Another concern is measurement accuracy. Most of the current automatic acoustic analyses have been developed for speech recognition and synthesis. Since these methods are often application-oriented (e.g. categorizing voiced and voiceless stops instead of survey the distribution of VOT), they in general have lower requirement for accuracy, which makes them less ideal for corpus phonetics studies.

In this paper, we attempt to fill this gap by presenting a methodological study on the automatic extraction of vowel formants from a speech corpus. The main procedure involves robust LPC (RLPC) analysis (Lee, 1988), which augments traditional LPC by iterative reweighting of the signal, and dynamic formant tracking (Talkin, 1987), which tracks speech formant frequencies using dynamic programming. The goal of the current study is twofold. First, we would like to showcase the use of RLPC on a large speech corpus. Compared with conventional LPC analysis, RLPC offers significant improvement by modeling the glottal source in a more sophisticated way. In this paper, we will discuss in detail the implementation, evaluation and optimization of RLPC on speech corpora. Second, it is also our goal to survey the distribution of vowel formants in spontaneous speech and compare with previous results from word production in isolation (Hillenbrand et al., 1995).

The remainder of this paper is organized as follows. The rest of this section is a brief introduction to the RLPC algorithm. Section 2 describes the dataset and the implementation and optimization of the formant analysis. Section 3 presents the measurement results based on the optimal formant analysis and shows the distribution of vowels across speakers. Section 4 concludes with a brief summary of the current study.

1.2. Robust LPC algorithm

Linear predictive coding (LPC) is currently the most widely used method for finding formants in the spectra. For voiced speech it assumes an autoregressive model, in which the sample at time n (i.e. x(n)) is a linear combination of previous samples (i.e. x(n-1), x(n-2), etc). One problem with the LPC analysis is that the autoregressive model assumes that the signal is stationary. In speech there are two sources of non-stationarity: the resonances of the vocal tract are highly damped, and with each pulse of the voice source new energy is added to the signal.

The RLPC method used in this paper (based on the Robust Linear Prediction algorithm in Lee [1988]) refines LPC by relaxing the assumption of signal stationarity. RLPC downweights the errors introduced by the large-variance impulses in order to provide a lower-variance estimate of the LP coefficients, resulting in better source/filter separation as well as lower-variance formant and bandwidth estimates. Section 2.2 below contains more technical detail about the implementation of the RLPC algorithm.

2. Methodology

2.1. Corpus and dataset

This study uses data from the Buckeye Corpus of Conversational Speech (Pitt et al, 2007), which contains 40 speaker's speech of about 300,000 words. All subjects were local residents of Columbus, Ohio, USA. Each subject was interviewed for about an hour for their opinion on various issues in life (without knowing the real purpose of the interview). The subjects were interviewed in one-to-one sessions by the experimenters. Only the interviewees' speech was digitally recorded and phonetically transcribed. Age and gender were balanced among the speakers: 20 speakers were male and 20 were female; 20 were old (>40 y.o.) and 20 were young (<= 40 y.o.). The youngest speakers were in their late teens and the oldest were in late seventies.

Speech recordings were transcribed both at the word level and at the phoneme level. The onsets and offsets of words and phonemes were labeled by a forced alignment algorithm and then hand corrected. Discourse fillers (e.g. *um* and *uh*) as well as nonlinguistic sounds (e.g. laughter and coughing) were also included in the transcription.

Figure 1 shows two clips of speaker S01's transcription files, one at the word level and the other at the phoneme level. As shown in Figure 1, Speaker S01 starts pronouncing the word "about" when t=345.722s (i.e. when the previous word *talk* ends) and the word ends at t=345.946s. Instead of pronouncing the word as "ah b aw t"[1] as in the dictionary, the speaker produces "b ah t", which consists of three phonemes, starting at t=345.722s, 345.785s, 345.887s, respectively.

```
Word-level transcription
End time    word        phonetic transcription
…
345.722     talk        t ao k
345.946     about       b ah t
…

Phoneme-level transcription
End time    phoneme
…
345.722     k
345.785     b
345.887     ah
345.946     t
…
```

Figure 1. Part of transcription files s0101a.word (upper) and s0101a.phone (lower), of speaker S01.

The current study focuses on the production of ten monophthong vowels ("aa", "ae", "ah", "eh", "ey", "ih", "iy", "ow", "uh", "uw"). Diphthongs "ay", "aw" and "oy" are excluded since their formant patterns are more variable both within the token and across tokens, but "ey" and "ow" are included because they are more often pronounced as monophthongs in American English.

To form the dataset, we first extracted from the corpus all instances of the target vowels (based on the transcription, which may or may not match dictionary pronunciations). Then we excluded vowel tokens that were shorter than 40ms for more reliable formant analysis (since we were using 30ms window size). The final dataset contains a total number of 287,223 vowel tokens. Table 1 summarizes the number of tokens of each vowel type.

Table 1 Average number of tokens per vowel type

aa	ae	ah	eh	ey	ih	iy	ow	uh	uw
14745	18654	70587	36380	15056	58064	33614	17645	6054	11423

2.2. Acoustic analysis

2.2.1. Signal preprocessing

The speech signal is first downsampled to twice the highest expected formant frequency (cf. section 2.2.3) and high-pass filtered with a cutoff at 80 Hz. It is also preemphasized with a first order difference equation as stated in (1). The preemphasis process will be adjusted in the assessment stage (cf. section 2.3.1).

(1) $y(n) = x(n) - c \times x(n-1)$, where c is the preemphasis factor

2.2.2. RLPC analysis

Our RLPC algorithm begins with a conventional LPC analysis (30 ms Gaussian window, 10 ms frame step, LPC order of 6, 8, or 10; cf. section 2.3) and continues with a method for refining the coefficients (we use the Iterative Weighted Least Squares method, as suggested in Lee [1988]). Lee's method first scales the residuals of the LPC analysis using a minimax estimator (Huber, 1964) , which decreases the weight of the small number of outlier residuals while leaving the much larger number of small to moderate residuals unchanged. Then the scaled residuals are used to recalculate the LP coefficients with a weighted least squares equation (Lee's equation 3.9, cited in (2)).

(2) $$\sum_{n=p+1}^{N} S_{n-j}\varepsilon_n(a^{(k+1)})W(\varepsilon_n(a^k)) = 0, \qquad 1 \le j \le p$$

In this equation S is the sampled data signal, p is the LP order, ε is the LP residuals, W is Huber's minimax estimating function, a is the autoregressive coefficients, and a^k indicates the k^{th} iteration of the solution. Hence, the current iteration's LP residuals are multiplied by the weighted residuals of the previous iteration. The system of equations defined by (2) is converted to matrix form and solved using standard linear algebra (see Lee [1988] for details).

In theory the refining method can be applied repeatedly, but in practice we found that the greatest improvement was in the first pass and using more than two iterations was not very useful. Results presented in the following are based on 2-iteration RLPC analysis.

2.2.3. Dynamic formant tracking

Dynamic formant tracking (Talkin, 1987) is used to find the most likely formant trajectories over time, based on the formant frequency candidates at each time step as identified by RLPC. The algorithm works by keeping track of the cost of different frequency-to-formant mappings and selecting the one with the lowest cost. Two types of cost are calculated for each mapping: a local cost and a transition cost. Local cost is based on the comparison between formant frequencies in local frames and a predefined expected frequencies matrix (we used expected normative values based on Hillenbrand et al, 1995, as well as expected minimal and maximal values). The better they match, the lower the cost is. Transition cost, on the other hand, is calculated for adjacent frames and penalizes frequency changes between frames. The two costs are combined to give an overall cost, and the frequency-to-formant mapping with the lowest overall cost will be selected (by a modified Viterbi algorithm) as the optimal formant trajectory.

2.3. Parameter optimization

The behavior of the above algorithms can be modified by varying a number of input parameters. First of all, the results of formant analysis can be influenced by preprocessing parameters, such as downsampling rate and preemphasis of higher-frequency signal components. In the RLPC analysis, the original LP coefficient estimates can be provided by either autocorrelation (the one we used) or covariance methods. One can also vary the order of the LPC models (in both the conventional analysis and later coefficients refinement), as well as the number of iterations that coefficient refinement will run. In the dynamic formant tracking algorithm, there are more parameters that can vary. Calculation of the mapping costs is dependent on the expected formant frequencies as well as the relative weights of different types of penalty (e.g. deviation from expected values, missing formants, formant merger, formant change from frame to frame, etc).

These parameters constitute a high-dimensional space. In principle, one can explore the entirety of the parameter space to determine the "optimal" parameters, but in practice, this is not feasible without a distributed computing grid. It was thus necessary for us to restrict the optimization to a parameter subspace in which one expects variation of parameters to have the largest impact on formant analyses.

In this study, we varied two parameters, preemphasis process and LPC order, to obtain optimal formant tracking results for each speaker and vowel. In the following, we will first propose a method for automatically evaluating the goodness of the analysis, and then present the optimization results of the two parameters.

2.3.1. Assessing formant analyses

Obviously the most reliable way to assess the success of a formant analysis is to check the measurement results against the spectrograms by hand, but this is not possible due to the large amount of data. In practice, we have observed that most of the errors in the formant analysis are due to missing formants (e.g. unable to find formant candidates when the signal is weak) and misidentification of formants (e.g. wrongly identifying H1 or H2 as F1 in high vowels). In view of this, we adopted an evaluation method which takes into account the missing formant rate (i.e. the percentage of frames with missing formants) and the variability of the formant measures. We expect a good formant analysis to be associated with a low percentage of missing formants and a relatively low variability in the formant measures (for the same vowel type produced by the same speaker). On top of that, it is also important that the analysis overall generate reasonable estimates of formant frequencies.

In practice, we have noticed that there is often a trade-off between missing formant rate and measurement consistency: when the missing rate is high (sometimes it can be as high as 50%), the formant measures tend to be less variable because fewer frames with inaccurate formant estimates are included in the calculation of variance. Hence we set an arbitrary standard for missing formant rate, that is, any analysis with more than 10% of frames with missing formants will be excluded from the running for the optimal analysis. The rest of the optimization process is solely based on variability of measured formants.

In order to quantify the variability of formant measures in both dimensions (i.e. F1 and F2), we employ a density area metric which is calculated as the area of F1,F2 space that contains 95% of the measurements from all analysis frames. The smaller this area is, the more convergent the formant measures are. Assuming that the formant measures are normally distributed, the 95% contour area of F1,F2 space in theory contains measurements that are within two standard deviations away from the mean. We refrain from using a 100% contour area in order to avoid the influence of outliers.

2.3.2. Varying analysis parameters

For each speaker/vowel combination, we conduct 3 (types of preemphasis) * 3 (LPC orders) = 9 formant analyses. We exclude analyses from consideration which produce more than 10% missing frames in either formant. We then determine the optimal analysis as the one with the smallest 95% density area in the F1-F2 space. The three versions of preemphasis filtering are the following: no preemphasis, 6dB/octave emphasis beginning at 500 Hz (which avoids pre-emphasis of high vowel F1) and 6 dB/octave emphasis beginning at 50 Hz (which pre-emphasizes the signal for all formants). The three LPC orders we use are 6, 8, and 10. According to the conventional wisdom, the order should be 2+2*NF, where NF is the number of expected formants in the range of frequencies represented in a signal. As mentioned above, the speech signal is downsampled to twice the maximal expected value of F2 for a given gender. Considering within-gender variation in vocal tract geometry as well as inter-token variation, NF should equal 2 or 3 in most situations (though it does equal 4 sometimes).

Figure 2 shows an example of how our optimatization metric varies as a function of LPC order and preemphasis for two vowels ("aa" and "iy") of two speakers (S04, female; S06, male). For speaker S04, the optimal analysis for the vowel "aa" is achieved when LPC order equals 6 and preemphasis starts at 500Hz, while for "iy", when LPC order is 6 and there is no preemphasis. On the other hand, for speaker M06, the optimal setting for vowel "aa" is when LPC order equals 10 and preemphasis starts at 50Hz, while for vowel "iy", when LPC order equals 6 and there is no preemphasis. It can be seen from Figure 2 that by using the optimal parameter setting, the 95% contour area can be reduced by about 20 to 30kHz^2.

Table 2 summarizes the number of subjects with the corresponding optimal parameter setting. An ordered (multinomial) logistic regression on optimal LPC order reveals significant effects of speaker gender (p=0.03) and vowel frontness (p<0.001). Generally speaking, male speakers tend to favor higher LPC orders than female speakers, while back vowels also favor higher LPC orders than front vowels. However, an ordered logistic regression on optimal preemphasis factor shows no effects of gender, vowel height or vowel frontness (p>0.2 in all cases).

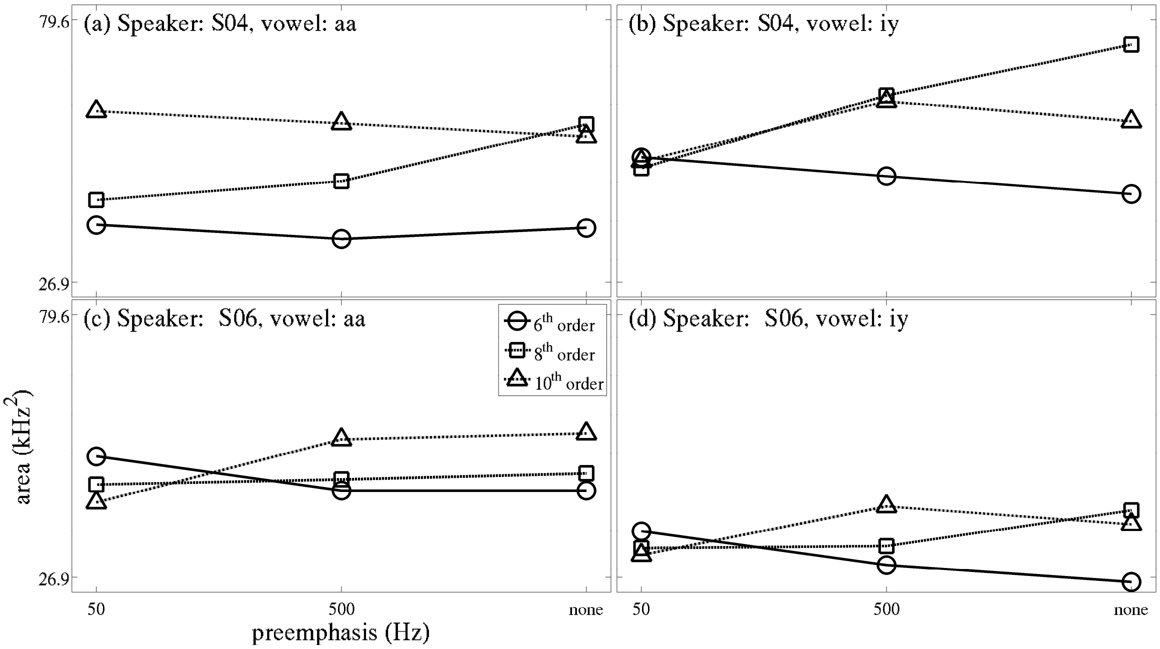


Figure 2. Parameter optimization based on 95% density areas (in F1-F2 space). The lines denote the optimization contours of different values of LPC order. X-axis represents preemphasis setting. Y-axis is the corresponding 95% density areas in kHz^2. Upper: Speaker S04's vowel "aa" (left) and "iy" (right); Lower: Speaker S06's vowel "aa" (left) and "iy" (right)

Table 2 Distribution of optimal preemphasis setting and LPC order among subjects

		Female speakers									
		aa	**ae**	**ah**	**eh**	**ey**	**ih**	**iy**	**ow**	**uh**	**uw**
Optimal preemphasis	**none**	9*	7	4	14	8	6	7	4	12	11
	50Hz	7	10	8	2	6	10	9	14	6	7
	500Hz	4	3	8	4	6	4	4	2	2	2
Optimal LPC order	**6**	14	2	0	5	16	7	15	2	5	5
	8	2	11	11	9	4	2	2	6	3	6
	10	4	7	9	6	0	11	3	12	12	9
		Male speakers									
		aa	**ae**	**ah**	**eh**	**ey**	**ih**	**iy**	**ow**	**uh**	**uw**
Optimal preemphasis	**none**	8	11	7	12	3	1	7	8	5	3
	50Hz	10	4	9	4	16	19	11	6	15	13
	500Hz	2	5	4	4	1	0	2	6	0	4
Optimal LPC order	**6**	8	6	5	4	7	7	11	3	0	2
	8	1	6	6	5	6	5	7	4	6	11
	10	11	8	9	11	7	8	2	13	14	7

* The values in the cells indicate the number of speakers with the corresponding optimal preemphasis and LPC order.

3. Results

3.1 Average durations and formant frequencies

Table 3 shows the average values of vowel duration and the first two formant frequencies, separated by gender and vowel type. The numbers represent the mean of individual speakers' mean values over all tokens. Figures 3(a) and 3(b) plot the average vowel space of each individual speaker, grouped by gender. For better resolution, the six tense vowels ("iy", "ey", "ae", "aa", "ow" and "uw") are plotted separately from the lax vowels ("ih", "eh", "ah" and "uh") on the vowel chart.

Table 3 Average duration (ms) and formant frequencies in men and women

		aa	**ae**	**ah**	**eh**	**Ey**	**ih**	**iy**	**ow**	**uh**	**Uw**
Dur (ms)	**w**	117	139	86	88	126	69	98	144	68	108
	m	109	131	87	86	121	68	91	130	67	102
F1 (Hz)	**w**	760	743	615	625	494	455	397	614	500	441
	m	644	629	548	535	451	432	350	522	456	399
F2 (Hz)	**w**	1342	1803	1442	1847	2296	1967	2408	1466	1556	975
	m	1255	1676	1350	1689	1857	1745	2182	1193	1426	867

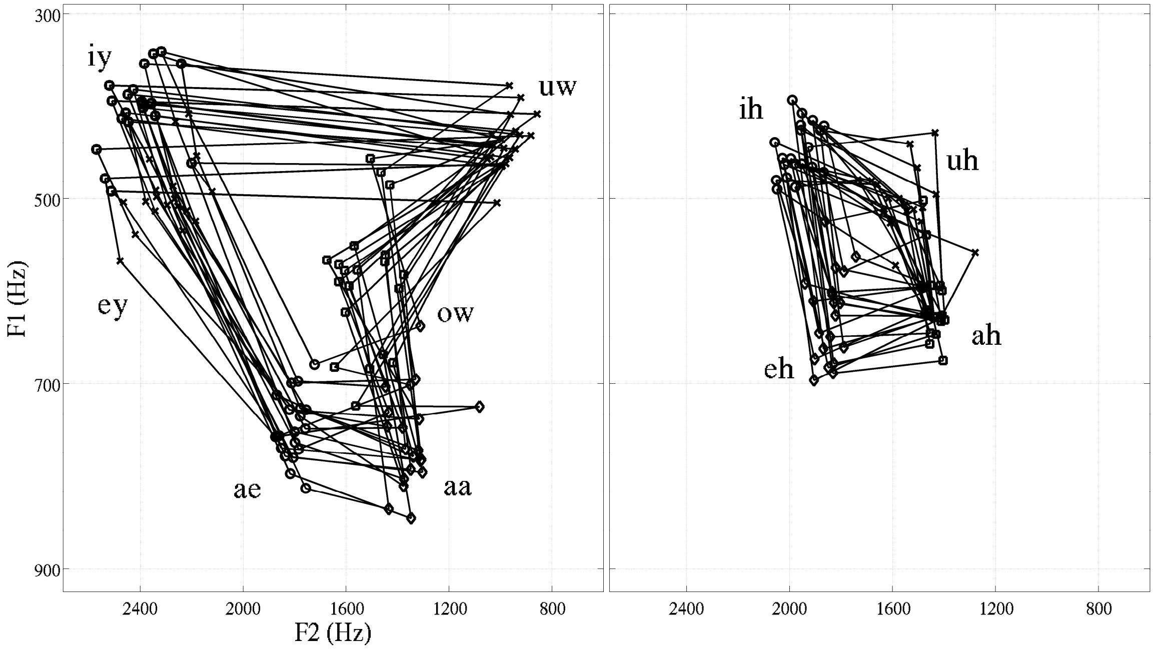


Figure 3(a). Average vowel formants of tense (left) and lax (right) vowels in female speakers. Each polygon represents an individual speaker.

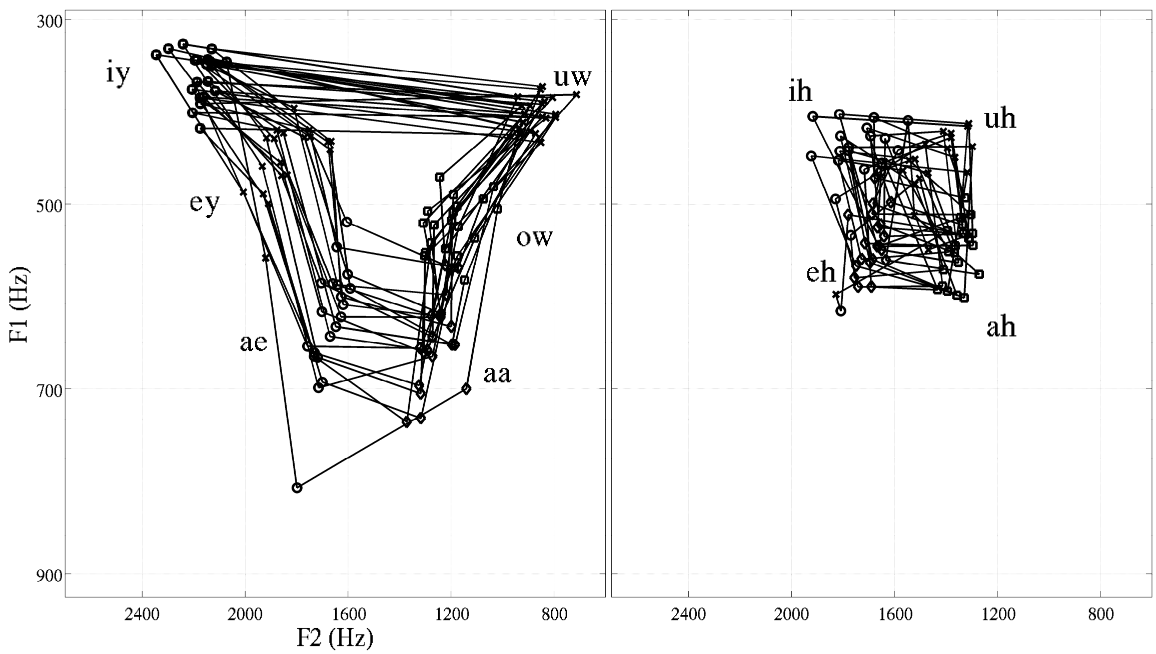


Figure 3(b). Average vowel formants of tense (left) and lax (right) vowels in male speakers. Each polygon represents an individual speaker.

3.2. Inter-speaker variation

Generally speaking, female speakers produce slightly longer vowels than male speakers (*p*=0.009). As expected, they also have higher formant frequencies compared to male speakers (*p*<0.001 for F1, *p*=0.002 for F2). More importantly, as can be seen from Figure 3, on average female speakers have a much larger vowel space than male speakers. This is also consistent with previous findings (Byrd, 1994). Both longer duration and more expanded vowel space are indicators of clear speech (Bradlow et al., 1996), which suggests that female speakers produce clearer speech than male speakers.

In both genders, individual differences in average formant values are not big – in most cases, the standard deviation across speakers is below 100 Hz (cf. Table 4).

Table 4 Inter-speaker differences in average formant frequencies in men and women

		aa	**ae**	**ah**	**eh**	**ey**	**ih**	**iy**	**ow**	**uh**	**uw**
s.d. in mean F1	**w**	53	35	39	45	41	31	47	55	37	31
(Hz)	**m**	54	60	35	29	48	33	20	23	39	18
s.d. in mean F2	**w**	76	40	30	44	100	56	96	87	107	57
(Hz)	**m**	60	57	49	45	127	122	76	91	133	60

3.3. Within-speaker variation

The current optimization method assumes that a good formant analysis (i.e. one with fewer dropped frames and less variability) should give a reasonably high degree of coherence when measuring the same vowel of the same speaker. However, how much of within-speaker variability we should expect is a tricky question. In fact, as mentioned in the introduction, within-speaker allophonic variation is one of the major research topics that speech corpora are used to investigate. In this exploratory work, we use 95% density area in F1/F2 space to measure formant consistency, where all analysis frames are taken into consideration. Based on the optimal parameter setting (which gives the smallest contour areas), the standard deviation in formant measures for each speaker/vowel combination is around 100 Hz in F1 and between 150 – 350 Hz in F2 (cf. Table 5).

Table 5 Average within-speaker variation in formant frequencies in men and women

		aa	**ae**	**ah**	**eh**	**ey**	**ih**	**iy**	**ow**	**uh**	**uw**
Mean s.d. in F1	**w**	95	111	115	102	82	79	66	140	72	67
(Hz)	**m**	84	91	91	85	90	118	57	76	137	58
Mean s.d. in F2	**w**	173	137	178	159	228	225	241	358	268	135
(Hz)	**m**	163	139	151	165	220	273	216	222	326	144

3.4. Comparing with single word production

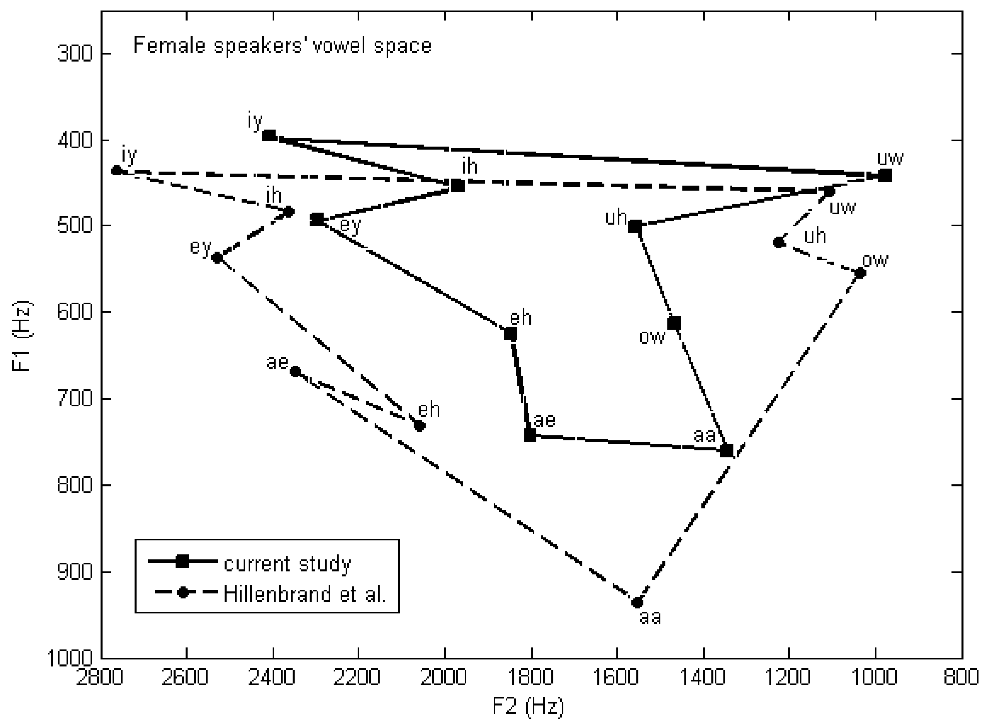
Hillenbrand et al (1995) reported vowel formants in men, women and children in single word reading. Their talkers consisted of 45 men, 48 women, and 46 children aged 10 to 12 (27 boys, 19 girls). Most of the talkers (87%) were from Michigan, and the remainder were from other areas in the midwest. Recordings were made of the subjects reading lists of /hVd/ words, with twelve different vowels. In addition to the ten vowels that are investigated in the current study, Hillenbrand et al. also recorded the vowel "ao" (as in the word *caught* when a *caught-cot* distinction is preserved) and the vowel "er" (as in *bird*). Hillenbrand et al.'s measurement provides reference points for later studies on vowel formant frequencies. It also forms the basis of the expected formant frequencies used in the formant tracking algorithm of the current study. In this section, we compare our measurement results with those in Hillenbrand et al (for men and women only; cf. Table 6).

Table 6 Average durations and F1/F2 formant values of ten vowels of men and women in word reading experiments in Hillenbrand et al's study

		aa	**ae**	**ah**	**eh**	**ey**	**ih**	**iy**	**ow**	**uh**	**uw**
Dur (ms)	**w**	323	332	226	254	320	237	306	326	249	303
	m	267	278	188	189	267	192	243	265	192	237
F1 (Hz)	**w**	936	669	753	731	536	483	437	555	519	459
	m	768	588	623	580	476	427	342	497	469	378
F2 (Hz)	**w**	1551	2349	1426	2058	2530	2365	2761	1035	1225	1105
	m	1333	1952	1200	1799	2089	2034	2322	910	1122	997

Compared with the current results, vowels in Hillenbrand et al.'s results are significantly longer (p<0.001 for both men and women), in fact, almost twice as long (mean vowel duration in the current study is 102 ms; mean vowel duration in Hillenbrand et al. is 259ms). Similar to the current results, there is a significant gender difference (p<0.001) in vowel duration, in that female speakers produce longer vowels (mean = 287ms) than male speakers (mean = 231ms).

On the other hand, formant measures in Hillenbrand et al.'s do not reliably differ from the current measures (p>0.1 for both F1 and F2, in both men and women), partly because we used Hillenbrand et al.'s measures to form the expected frequencies matrix in dynamic formant tracking. However, as shown in Figure 4, the current measurement does reveal a less expanded vowel space for both men and women, compared with Hillenbrand et al.'s results. This is not surprising given the great difference in vowel duration, since we know that longer vowels tend to have more extreme formant values than shorter vowels ("duration-dependent vowel undershoot/overshoot"; Moon and Lindblom, 1994). Together, the differences in both duration and vowel space expansion are consistent with the general consensus that isolated word production is featured by hyperarticulation whereas spontaneous speech contains a lot more phonetic reduction. We also speculate that the current results contain wider ranges of within-speaker variation than earlier results, because of the nature of the speech task as well as the variability of phonetic context in spontaneous speech. But a direct comparison is not available due to the lack of information about within-speaker variation in Hillenbrand et al.



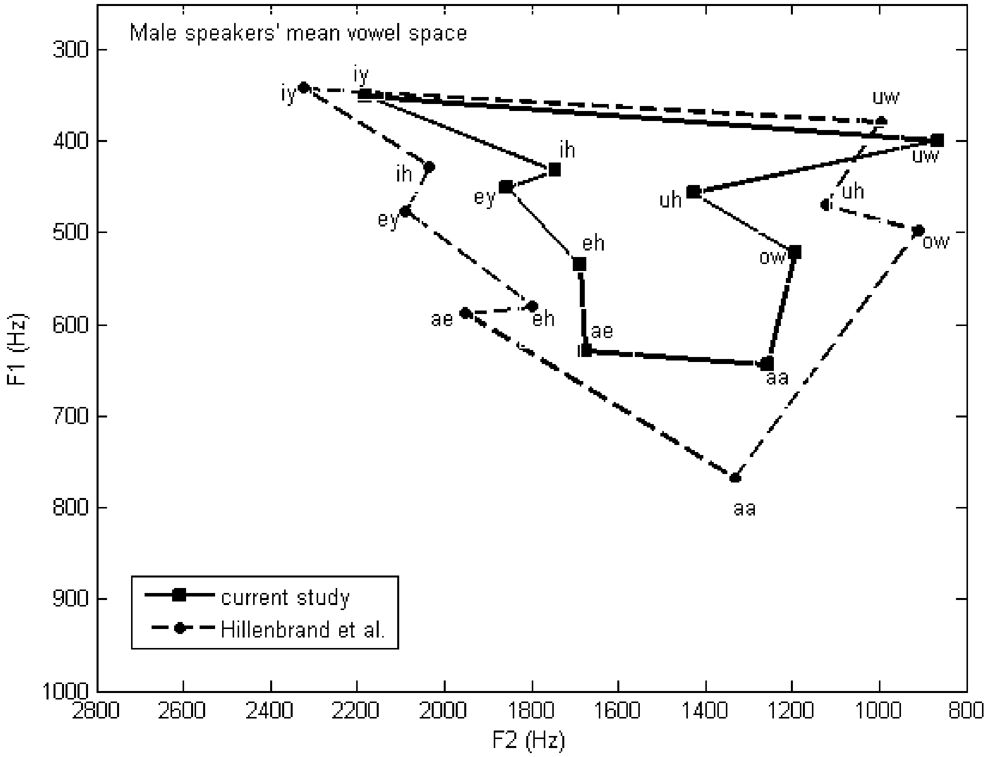


Figure 4. Comparing average values of F1 and F2 in the current study and in Hillenbrand et al.

4. Conclusion

To conclude, in this paper, we present an exploratory study on automatic vowel formant analysis using data from speech corpora. We use both robust LPC and dynamic formant tracking to automatically locate vowel formants in transcribed speech. We also explore the optimization of the automatic analysis by varying two parameters, the preemphasis process and the order of the LPC models. The resulting formant measures are consistent with previous findings on vowel targets in American English as well as gender difference in speech production. In our future work, we plan to further this study by investigating in more depth the evaluation and optimization of the current formant analysis.

Our long term goal in this line of research is to develop an automatic formant analysis that remains accurate and effective in presence of the variability in natural speech. Such an analysis will achieve reliable formant measures across speakers, vowels and contexts, and provide valuable data for research on individual differences, phonetic variation and coarticulation. In addition, for each vowel token, it measures the formant trajectories over time, which goes beyond the average (or midpoint) values and allows researchers to investigate the time course of the phonetic processes.

Notes

[1] The Buckeye corpus uses DARPA phonetic alphabet in the transcription. In accordance, the same alphabet is used in this paper to refer to sounds. See the appendix for a table with both DARPA symbols and the corresponding IPA symbols.

References

Bell, Alan, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. (2009) Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1): 92-111.

Bell, Alan, Dan Jurafsky, Eric Fosler-Lussier, Cynthia Girand and Daniel Gildea. (1999) Forms of English function words - Effects of disfluencies, turn position, age and sex, and predictability. In *Proceedings of ICPhS-99*, 395-398, San Francisco, CA.

Bradlow, Ann R., Gina Torretta and David B. Pisoni. (1996) Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20: 255-272.

Byrd, Dani. (1994) Relations of sex and dialect to reduction. *Speech Communication* 15(1): 39-54.

Talkin, David. (1987) Speech formant trajectory estimation using dynamic programming with modulated transition costs. *Journal of Acoustical Society of America* 82(S1): S55-S55.

Fosler-Lussier, Eric and Nelson Morgan. (1999) Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication* 29(2): 137-158.

Hillenbrand, James, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. (1995) Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97(5): 3099-3111.

Huber, Peter J. (1964) Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35(1): 73-101.

Jacewicz, Ewa, Robert A. Fox, Caitlin O'Neill and Joseph Salmons. (2009) Articulation rate across dialect, age, and gender. *Language Variation and Change* 21: 233-256.

Keating, Patricia A., Dani Byrd, Edward Flemming and Yuichi Todaka. (1994) Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication* 14(2): 131-142.

Lee, Chin-Hui. (1988) On robust linear prediction of speech. *IEEE Transactions on Acoustics, Speech and Signal Processing* 36: 642-650.

Moon, Seung-Jae and Björn Lindblom. (1994) Interaction between duration, context, and speaking style in English stressed vowels. *Journal of Acoustical Society of America* 96(1): 40-55.

Pitt, Mark, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. (2007) Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu]. Department of Psychology, Ohio State University.

Raymond, William D., Robin Dautricourt and Elizabeth Hume. (2006) Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18: 55-97.

Shriberg, Elizabeth. (2001) To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31(1): 153-169.

Yuan, Jiahong, Mark Liberman and Christopher Cieri. (2006) Towards an integrated understanding of speaking rate in conversation. *In Interspeech 2006*, 541-544, Pittsburgh, PA.

Author's contact information:
Linguistics Department
University of California, Berkeley
Yao Yao
Berkeley, CA 94720, USA
e-mail: yaoyao@berkeley.edu

Appendix

Symbols for vowels in Darpa Phonetic Alphabet

Darpa alphabet	IPA symbol	Example
aa	ɑ	cot
ae	æ	bat
ah	ə	but
eh	ɛ	bet
ey	e	bait
ih	ɪ	bit
iy	i	beat
ow	o	boat
uh	ʊ	put
uw	u	boot
ay	ɑj	bite
aw	ɑw	now
oy	ɔj	boy
ao	ɔ	bought
er	ɚ	bird