Automatic formant extraction for sociolinguistic analysis of large corpora

Keelan Evanini, Stephen Isard, Mark Liberman

University of Pennsylvania, Philadelphia, PA, USA

keelan2@ling.upenn.edu, stepheni@seas.upenn.edu, myl@cis.upenn.edu

Abstract

In this paper, we propose a method of formant prediction from pole and bandwidth data, and apply this method to automatically extract F1 and F2 values from a corpus of regional dialect variation in North America that contains 134,000 manual formant measurements. These predicted formants are shown to increase performance over the default formant values from a popular speech analysis package. Finally, we demonstrate that sociolinguistic analysis based on vowel formant data can be conducted reliably using the automatically predicted values, and we argue that sociolinguists should begin to use this methodology in order to be able to analyze larger amounts of data efficiently.
Index Terms: formant prediction, corpus analysis, sociolinguistics

1. Introduction

Despite the recent increase in the number and availability of large speech corpara, the field of sociolinguistics has yet to benefit substantially from the wealth of potential acoustic data. Most sociolinguists hold the belief that automated acoustic analysis is too error-prone to produce consistent and accurate results [1], and thus avoid large-scale corpus research in general. Since sociophonetic analysis is still based primarily on manual measurements and annotations, the vast majority of such studies either use datasets that are too small to generalize reliably to a wider population or take a long time to be analyzed.

This paper considers one of the most common types of acoustic analyses used in sociolinguistic research: vowel formant extraction. Ever since the first instrumental sociolinguistic analyses of sound change in progress [2], sociolinguistic studies of vowel variation have focused almost exclusively on F1 and F2. These two values are claimed to capture the "most salient regional and social differences in the pronunciation of the vowels of North American English" [1] (although see [3] for an argument that other sources acoustic information should also be included in sociolinguistic analyses). Despite the fact that other representations, such as MFCCs, are commonly used in ASR tasks, formants continue to prove useful to phoneticians because of their low dimensionality, their correspondance to articulatory gestures, their resistance to transmission channel effects, and their ability to characterize linguistically relevant vowel distinctions.

Vowel formant extraction, however, as conducted in most sociophonetic analyses, is a laborious process which requires the annotator to listen to every token while examing a spectrogram before recording the F1 and F2 values from the LPC analysis program (although see [4] for a notable early exception). Often, the annotator adjusts the number of poles in the LPC analysis when formants are not visible in locations that would be expected based on the annotator's prior knowledge of the distribution of the vowel in the token. Such adjustments are frequently necessary, especially for non-high back vowels where the F1 and F2 values are close together (for example, based on the annotations in the log files for [1], this process was necessary in at least 10% of 134,000 manual formant measurements).

Due to the time required to produce sociolinguistic analyses of large speech corpora using the accepted techniques of manual formant analysis, it would be beneficial to the field if it could be shown that automatic formant analysis can produce results that reliably capture sociolinguistic variation. For example, the ANAE corpus (described in Section 2) took several years to analyze manually, whereas approximately 10 times as much data with similar sociolinguistic attributes could be automatically extracted from the Switchboard corpus [5]. This paper presents a method for automatic formant extraction for corpora where a transcript of the audio file is available, and thus the identity of the vowel to be analyzed is known. The general approach is to use the same type of knowledge as the human annotator does when making decisions in the manual formant extraction process: namely, prior knowledge of the distribution of formants for a given vowel. We will show that this method of formant extraction leads to an improvement over the formants produced by the default settings of a commonly used formant tracking program, and that the automatically extracted formants are able to accurately characterize groups of speakers based on their participation in regional sound changes.

2. Description of the Corpus

The sociolinguistic corpus used for the current analysis is the Atlas of North American English [1], henceforth ANAE. The ANAE corpus consists of ca. 30-minute long dialectological interviews conducted over the telephone with speakers from across the United States and Canada, and represents the most complete corpus of geographical variation in English. The ANAE sampling methods ensured that the data from each dialect region is more accurate and fine-grained than in other corpora: at least two speakers were selected randomly from every city in North America with more than 50,000 inhabitants, and only speakers who had lived their entire lives in that city were chosen. The interviews consisted of a series of elicitation tasks which focused on the pronunciation of specific lexical items, minimal pair tests, free conversation, and a word list.

A total of 439 speakers were selected for detailed acoustic analysis by the ANAE authors. For these speakers, annotators examined all tokens with primary stress, and provided hand measurements for the first two formants at a single point in time. The measurement points were chosen to represent the "central tendency" of each vowel, and were determined through a combination of auditory perception and visual analysis of the spectrogram (see Section 5.5 in [1] for a detailed description of the procedure). For tokens where the formants produced by the LPC software did not match the spectrograms, the annotators

modified the number of poles used in the LPC analysis until an acceptable formant track was produced. In total, 134,000 vowel were analyzed in this manner.

For the purposes of comparing automatic formant prediction methods with the F1 and F2 values provided by the human ANAE annotators, it is necessary to determine the point in time at which the manual F1 and F2 measurements are taken. This information is not contained in the log files that were included in the published version of ANAE, but is available in earlier versions of the log files obtained from the ANAE authors. These two sources of information were merged to produce a database of F1 and F2 measurements with time stamps for a total of 111,810 tokens from 384 speakers (formant data from several speakers had to be excluded because the original log files with time stamps were not available).

3. Methods

3.1. ESPS Formant Extraction

A baseline set of automatic formant measurements was extracted by using the `formant` command from the ESPS software package [6]. Most default settings for the `formant` command were used, resulting in the following formant analysis parameters: 12 order autocorrelation LPC analysis using a 49 msec raised cosine window at 100 Hz with a preemphasis factor of 0.7. The only setting given a different value was the number of formants to predict: this was set to 3, since the corpus consists of telephone speech and the signal thus has a maximum frequency component of 3500 Hz (tests conducted with the default setting of 4 formants resulted in similar performance). After the `formant` command was run on each token, the predicted F1 and F2 values at the point in time closest to the hand measurement were extracted.

3.2. Proposed Formant Prediction Method

The general approach taken by the proposed formant prediction method is to simulate the procedure used by a human annotator by incorporating prior knowledge of the distribution of formant and bandwidth combinations for specific vowels. For each vowel, a model of formant and bandwidth combinations was trained by computing the means and full covariance matrices for the manual F1 and F2 measurements with their respective bandwidths. Since bandwidth information is not provided in the ANAE corpus, the bandwidth values associated with the default ESPS formant tracker were used when they were close to the hand formants. The threshold we used for determining whether to use a token's ESPS bandwidth data in the training set was if both the predicted F1 and F2 values were within 7% of the respective hand measurements. This criterion led to a total of 61,048 training tokens (55% of the total corpus) with manual F1 and F2 values plus bandwidth data from the ESPS measurements (tests were also conducted with models trained using only F1 and F2 data from all 111,810 tokens, i.e. without bandwidth data, but this led to decreased performance). Additionally, the bandwidth measurements were converted to the log domain for both training and testing in order to make the bandwidth distributions closer to Gaussian (tests with the formant frequency values also converted to the log domain showed no further improvement).

To predict F1 and F2 using the current method for a given test vowel, we consider all possible pairs of poles and their associated bandwidths returned by the ESPS LPC analysis for the vowel. This results in $\binom{n}{2}$ test instances, where n is the number of poles provided by ESPS. Each test instance, x, is thus a vector consisting of four values: the two potential formant values and their associated bandwidths. To determine the most likely F1 and F2 values, the Mahalanobis distance, D, between the model for the vowel and each test instance is computed, and the two poles from the vector for which the distance is smallest are assigned to F1 and F2. The equation for D is given in Equation 1, where μ and Σ are the means and covariance matrix for the formant and bandwidth values for the vowel.

$$D(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (1)$$

4. Results

4.1. Comparison to Baseline

Table 1 shows the overall improvement over the ESPS formants using our predicted formants for all 111,810 tokens in the corpus, obtained by applying 10-fold cross validition to the entire data set (see [7] for similar results comparing ESPS and manual formants on a different database). The proposed method reduces the global mean absolute difference from the hand measurements by 10% in F1 and 20% in F2. The performance improved for all 30 individual vowel classes in F1; in F2, a large performance gain was obtained for non-front vowels, whereas performance declined slightly for front non-low vowels.

	Mean abs. difference		RMS difference	
	F1	F2	F1	F2
ESPS default	54.8	112.8	97.9	297.3
Proposed method	49.4	90.6	79.8	199.9

Table 1: Differences between two formant prediction methods and manual measurements (N = 111,810)

The observed mean differences between our automated measurements and the ANAE hand measurements are comparable in size to the generally-acknowledged uncertainty in formant frequency estimation (cf. also the inter-labeler agreement results reported in [7]), and also to the perceptual difference limens found by [8].

Figures 1 and 2 graphically illustrate the comparison between the proposed method and the baseline for 17,954 tokens from three vowel classes: IY as in *heat*, AA as in *hot*, and UW as in *hoot* (the UW class corresponds to the /Kuw/ class in ANAE, namely /uw/ after non-coronal onsets, where most dialects do not have substantial fronting). The most striking difference between the two sets of predicted formant values is in the lower-left quadrant of the two plots: in the ESPS plot in Figure 1 there are many tokens of AA and UW erroneously predicted to be in this quadrant, whereas the plot from the proposed method looks much more similar to the distributions obtained by the hand measurements shown in Figure 3.

4.2. Vowel classification

As an additional method of comparison between the different sets of extracted formants, three-way vowel classifications tasks were conducted. For these tests, all vowel formant measurements were normalized using the procedure in [9], which is the standard method used in sociolinguistic studies to reduce effects due to vocal tract length differences but still preserve effects due to sociolinguistic characteristics of the speaker [1]. Separate group means and speaker normalization factors were computed for each of the three different sets of F1 and F2 measurements.

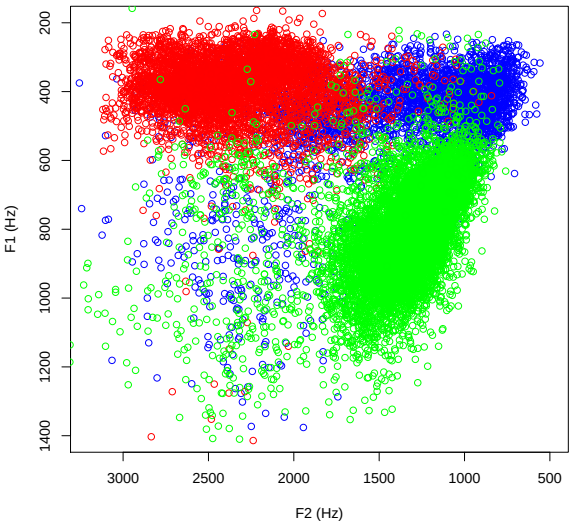

Figure 1: ESPS F1 and F2 measurements for IY, UW and AA

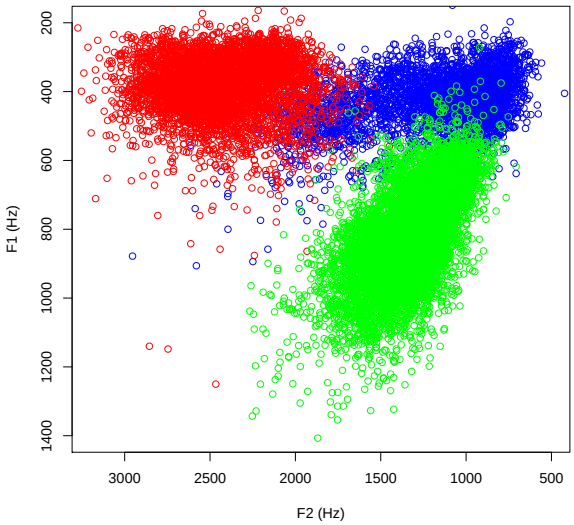

Figure 2: F1 and F2 measurements for IY, UW and AA predicted by the current method

Formant values	Accuracy
Manual	97.9%
Predicted (proposed method)	97.6%
ESPS default	90.2%

Table 2: Overall accuracy for classifying IY, UW, and AA using three different sets of F1 and F2 values (N = 17,954)

classified as →	AA	IY	UW
AA	9665	2	139
IY	5	4604	153
UW	32	103	3251

Table 4: Classification results using predicted formants

For the normalized versions of the 17,954 tokens of the three vowels AA, IY, and UW shown in Figures 1 – 3, the classification task was to predict the most likely vowel class given F1 and F2. Again, models for each vowel using the means and covariance matrices were trained through 10-fold cross validation, and classification was done for each test vowel by minimizing the Mahalanobis distance (similar results were obtained for all tasks using a decision tree classifier). Table 2 presents the overall classification accuracy for the three different sets of formant values, and Tables 3 – 5 show the confusion matrices.

The overall accuracy and vowel-specific precision and recall using the F1 and F2 values predicted by the proposed method are higher than the results using the formants predicted by ESPS, and are quite similar to the results obtained from using the hand measured formants.

4.3. Characterization of sound changes

The most important test for the applicability of any automatic formant prediction method from a sociolinguist's perspective is whether the predicted values demonstrate the same group trends as the values measured by hand. I.e., for any sociolinguistically important group of speakers based on sex, age, geographic region, etc., the group's vowel means from the predicted formants must demonstrate the same type of variation as the means from the hand measurements. Even when a small number of automatic formant measurements are gross errors, the sociolinguistic analysis can still be conducted successfully if they are not systematically biased in any direction.

As a demonstration of this approach to validating the automatic formants predicted for the ANAE, Figure 4 displays the vowels participating in the Northern Cities Shift (NCS). This large-scale chain shift involves most of the short vowels, and is currently proceeding in a strikingly uniform manner throughout most of the Inland North dialect region (see Chapter 14 in [1]). As a reference for the non-shifted forms, Figure 4 shows the means (from non-prenasal tokens) from the normalized manual F1 and F2 measurements for the 332 non-Inland North speakers in black. The NCS vowel means for the 52 Inland North speakers are then shown for the three sets of measurements.

The two sets of automatically predicted formants both capture the characteristics of the NCS: tense AE is higher and fronter than EH, which has lowered and moved back; also the NCS strong fronting of AA is clearly visible. While none of the NCS vowel means predicted by either of the two automatic methods are far enough off from the hand measurements to obscure the relative positions of the vowels, the values predicted by the current method for the vowels AH and AO are much closer to the

classified as →	AA	IY	UW
AA	9726	1	79
IY	7	4594	161
UW	33	100	3254

Table 3: Classification results using hand formants

classified as →	AA	IY	UW
AA	9033	11	762
IY	17	4130	615
UW	109	250	3027

Table 5: Classification results using ESPS formants

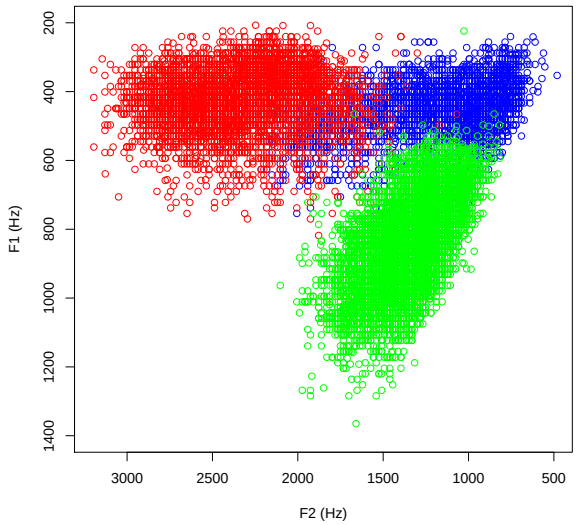

Figure 3: Hand F1 and F2 measurements for IY, UW and AA

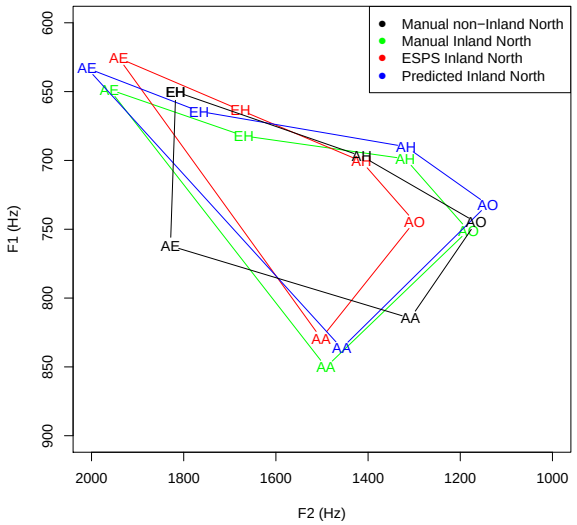

Figure 4: Northern Cities Shift vowel means for Inland North and non-Inland North speakers

hand measured means than the ESPS values, while the ESPS mean for EH is closer than the one predicted by our method. In either case, a sociolinguistic analysis using the automatically predicted formants would reach the same conclusions as one using the manual formants.

5. Discussion

The analysis of the Northern Cities Shift vowels in Section 4.3 shows convincingly, in our opinion, that sociolinguistic research can be carried out reliably by using automatically predicted formant values. We believe that the common stance of sociolinguists that only manual formant values are accurate enough for sociophonetic analysis is no longer valid. For corpora which have transcriptions available, and can thus be subject to forced alignment so that vowel boundaries can be obtained, automatic formant values can lead to reliable sociolinguistic analyses much faster than the traditional method. Furthermore, when very large corpora are used, errors in individual tokens and even individual speakers will not harm the analysis.

Additionally, the overall differences in Section 4.1 and the classification results in Section 4.2 show that our method of predicting formants by using prior knowledge of the formant distribution for the vowel in the reference transcription performs better than the default settings of a popular formant tracking program. Further refinements to this Bayesian approach to formant prediction, especially the inclusion of robust statistics to identify errors, could reduce the difference between manual formant measurements and automatic ones even further.

The final step that is necessary in producing a fully automatic formant prediction system is to determine where in the vowel the formant measurements should be taken. For monophthongal vowels, this choice is not difficult, and nearly all time points outside of the consonantal transitions will produce acceptable measurements. For vowels with more complex trajectories, however, common approaches such as measuring a few fixed points or taking averages obscure the true nature of the vowel's target, especially in cases of sound change. Future work will attempt to address this question by using the manual measurements from the ANAE corpus to provide models (perhaps dialect-specific ones in cases of sound change) for the optimal measurement point for each vowel. Finally, further research will investigate reducing the size of the training set in order to determine the minimum amount of training data necessary to achieve reliable formant predictions for each vowel.

6. Conclusions

In this paper we have demonstrated that automatic formant tracking is a tool that is ready to be added to the methods that sociolinguists use to analyze acoustic data. The field would benefit substantially by adopting this methodology, due to the recent drastic increase in large corpora with sociolinguistically interesting distributions of speakers.

7. References

[1] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English*. Mouton de Gruyter, 2006.

[2] W. Labov, M. Yaeger, and R. Steiner, *A quantitative study of sound change in progress*. Philadelphia: U.S. Regional Survey, 1972.

[3] E. R. Thomas, "Applying phonetic methods to language variation," *American Speech*, vol. 75, no. 4, pp. 368–370, 2000.

[4] M. Lennig, "Acoustic measurement of linguistic change: the modern Paris vowel system," Ph.D. dissertation, University of Pennsylvania, 1978.

[5] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," *Proc. ICASSP*, pp. 517–520, 1992.

[6] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *Journal of the Acoustical Society of America*, vol. 82, no. S1, p. S55, 1987.

[7] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Cheng, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," *Proc. ICASSP*, 2006.

[8] P. Mermelstein, "Difference limens for formant frequencies of steady-state and consonant-bound vowels," *J. Acoust. Soc. Am.*, vol. 63, no. 2, pp. 572–580, 1978.

[9] T. Nearey, "Phonetic feature system for vowels," Ph.D. dissertation, University of Connecticut, 1977.