IMPROVED FORMANT FREQUENCY MEASUREMENTS OF SHORT SEGMENTS

David J.M. Weenink

University of Amsterdam
David.Weenink@uva.nl

ABSTRACT

We describe an algorithm that automatically finds the smoothest formant trajectories for short segments of speech. The method selects for each segment the smoothest from a number of alternatives. The smoothness criterion is based on the modeling of formant tracks with polynomial functions and uses both the χ^2 badness-of-fit as well as the variances of the polynomial coefficients. A great advantage with respect to other methods is that it is completely automatic and reproducible because of our new criterion that *quantifies* the smoothness of formant tracks. Applied to some speech corpora, the new method shows smaller spreading ellipses especially for male's high back vowels.

Keywords: formant frequency measurement, variable ceiling, polynomial approximation, smoothness

1. INTRODUCTION

Formant frequency measurements are used in linguistics as part of the acoustic description of vowel-like sounds. However, formant frequencies are notoriously difficult to measure in a fully automatic way. Nowadays most of the time they are determined with the help of linear predictive coding (lpc) algorithms and manual interventions. The basis of lpc analysis of a speech signal is the source-filter model of speech production. According to this model, a sound is described as being the results of passing a source signal through a filter that consists of a fixed number of (formant) filters. In lpc analysis the number of filters has to be chosen beforehand and good results are obtained only if the number of formants present in the signal matches the number of formants requested by the analysis software. If, however, the number of actual formants in the speech signal does not match the number of filters in the analysis, the results of the analysis can be poor. If the number of filters is chosen too large, i.e. larger than the number of actual formants, spurious formant tracks will appear and if, on the other hand, the number of filters is chosen too low, then some formants might not be measured at all, or, two or more formants might be averaged which results in one formant at the wrong frequency. These deficits of lpc are known since its incubation [6]. Because of this deficit, the actual number of formants in the sound segment and the requested number of formants of the analysis software may not match and therefore automatic error-free measurements of formant frequencies remains an illusion.

Often a lot of post-processing has to be done after the automatic measurements to "correct" those segments where the estimated formant frequencies were poor. In [4] the authors describe their tedious work to measure the acoustic characteristics of American English vowels with an interactive analysis based on lpc analysis. They also note that frequently decisions on the correct formant frequency values were influenced by the experimenter's knowledge of acoustic phonetics like the knowledge of the close proximity of the first two formants for /ɔ/ and /u/, or the proximity of F_2 and F_3 for /i/. More recently, in an acoustic description of the vowels of Northern and Southern Dutch [1] the authors estimate that in 20-25% of their cases they had to manually modify automatically obtained formant frequency tracks. Different kinds of ad hoc procedures have been developed to correctly measure formant frequencies. These interactive procedures make formant frequency measurements costly in terms of experimenter time and, worst of all, very difficult to reproduce. An example of such an advanced interactive method to determine formant frequency values is the one by Nearey et al.[7] which was used by Adank et al.[1]. In the Nearey et al. procedure a number of different lpc analyses with variable number of coefficients and formant ceilings are calculated and the experimenter chooses between them by ranking a number of criteria.

In [3] an optimal ceiling strategy was used to measure formant frequency values. Analyses were performed by applying lpc analysis on the same sound by varying the sound's bandwidth. After all analyses are performed, a speaker and vowel dependent variance minimization is performed, based on multiple reproductions of the same vowel by the same speaker, and the optimal bandwidth of the sound

(ceiling) and its formant frequencies were determined. This procedure resulted in a considerable reduction in the spread of formant frequency values.

Both methods have in common that they don't use a fixed number of coefficients and neither a fixed bandwidth of the sound in their analyses. The Neary et al. method also emphasizes continuity and smoothness of formant tracks. In the current paper we describe a new method that combines aspects of the two methods mentioned above. The greatest advantage with respect to these previously mentioned methods is that it is completely automatic and reproducible because we have developed a new measure that *quantifies* the smoothness of formant tracks.

2. THE NEW METHOD

Our method is optimized for small vowel-like speech segments with durations typically that of monophthongs or diphthongs, as most of formant frequency analysis reports in the linguistic literature are based on segments like these. The method consists of two steps.

In the first step we perform multiple lpc analyses on the same speech sound, this can be with a varying number of prediction coefficients or with varying sound bandwidths (ceilings). These analyses are all performed in a standard way. We start by downsampling to the desired bandwidth. We go on with performing an lpc analysis for overlapping frames of 25 ms duration. The prediction coefficients are calculated and transformed to formant frequency and bandwidth pairs. The latter values are stored in a so called Formant object. After the analyses we have a number of Formant objects, each Formant object contains, at regularly spaced time steps, the analysis frames with five formant frequency and bandwidth pairs. We can construct formant tracks by connecting corresponding formant frequency points from succeeding analysis frames. A formant track shows how a particular formant changes as a function of time.

The second step models in each Formant the formant tracks with polynomial functions of time. We calculate a smoothness value for each formant track and combine the values for all the tracks to one smoothness value. The Formant object with the best smoothness value is then chosen as the best representation of this segment.

Now we will explain in somewhat more detail the modeling process and how the smoothness criterion is quantified. In a vowel-like speech segment we model each formant track with a sum of polynomial functions. For our purpose Legendre polynomials are convenient functions as they are orthogonal. Legendre polynomials were also used by [10] to model formant movements in Dutch texts read at normal and fast rate. The model for a formant track $\hat{f}(t)$ is as follows

$$(1) \quad \hat{f}(t) = \sum_{k=1}^{m} a_k L_k(t),$$

where m is the number of Legendre polynomials to include in the model, $L_k(t)$ is the k^{th} Legendre polynomial, and a_k is the weight of the k^{th} Legendre term. The coefficients a_k have to be determined from the actual formant data. Because the domain of Legendre polynomials is bounded to the interval $[-1,1]$, we have to scale times appropriately. The number of polynomials m that we actually use in the model is very low, and, for vowel-like segments a value of three or four for m will normally be sufficient.

The coefficients a_k for a formant track are calculated in the standard way from the measured formant frequencies and bandwidths in a track by minimizing a χ^2 badness-of-fit criterion defined as follows:

$$(2) \quad \chi^2 = \sum_{i=1}^{n} \left(\frac{f_i - \hat{f}_i}{b_i} \right)^2 .$$

Here n is the number of *measured* formant frequencies in this track. If a particular formant frequency is not present in an analysis frame, we simply don't include its value in the estimation. Therefore the value of n is always smaller than or equal to the number of analysis frames in the segment. The f_i is the measured formant frequency at the i^{th} time position in the segment, b_i is the bandwidth of this formant and $\hat{f}_i$ is the modelled frequency value at the i^{th} time position as calculated from formula (1). We include the bandwidths as an indicator of how certain we are of the measured values f_i because we want to give formants that have sharp peaks in the spectrum (small bandwidths) more weight than formants that correspond to broad peaks (large bandwidths).

The system of equations that follows from minimizing expression (2) with respect to the a_k can be elegantly solved for the a_k by singular value decomposition [8]. Besides giving the solution that minimizes (2), singular value decomposition also gives the variances and covariances of the estimated parameters a_k. We have now modelled each formant track in a Formant object separately by equation (1) by minimizing its badness-of-fit value defined by (2).

From the χ^2 badness-of-fit function (2) it shows that the closer the estimated values $\hat{f}_i$ are to the measured values f_i, the smaller the value of χ^2. A small

value for χ^2 may indicate that the measured values lie "close" to the model function, whose trajectory is a smooth curve, and, therefore, that the measured data are "smooth" too. In this respect modeling is essential and superior to smoothing or formant tracking methods because it gives us an indication of the smoothness of the data on an *interval*. All discontinuities in the data are registered in the χ^2 value of a formant track. The total badness-of-fit χ^2 value of the complete model of the segment, i.e. all modelled formant tracks combined, can now simply be calculated by summing the χ^2 values from the individual formant tracks in the Formant object.

To select the best model for the segments it does not suffice, however, to pick the Formant object that has lowest combined χ^2 value. Formula (2) shows a tendency to favour models that have large bandwidths. For example, if we multiply all bandwidths b_i by a factor two, the new χ^2 value is reduced by a factor four: we got a better fit simply by increasing all bandwidths! This is not as we want it to be, because we would like to have the tracks with the *smallest* bandwidths to be selected as the best. Luckily these larger bandwidths translate directly into a larger variance of the estimated parameters. In the example above, if we double the bandwidths, the variance of each estimated parameter a_k also quadruples. This suggests that we also have to include the variance of the parameters in the selection criterion. We obtain a usable criterion to weigh complete models against each other if we use

$$(3) \quad W = \left(\frac{s_p^2}{k}\right)^t \left(\frac{\chi^2}{d}\right)$$

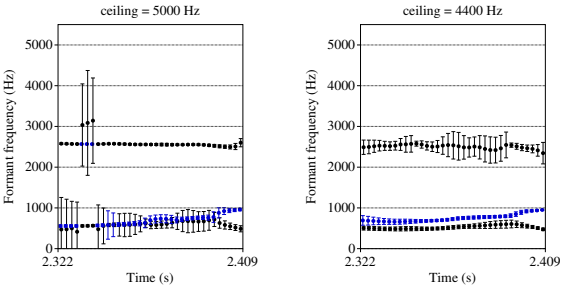
as the overall smoothness criterion, where s_p^2 is the sum of the variances of all the parameters of all modelled formant tracks within one Formant object, k is the sum of the number of parameters of these tracks, χ^2 is the combined value of these tracks, d is the combined number of degrees of freedom of these tracks, and t is a number that raises the combined variance to some power. The expression above grows like "bandwidth" to the power $2t-2$ because s^2 is proportional to bandwidth squared and χ^2 is inversely related to bandwidth squared. Choosing t somewhat larger than 1 guarantees that for two tracks that only differ in their bandwidths, the track with the larger bandwidths obtains a larger value for the criterion value W. The best fit is therefore the one with the lowest value for W.

To summarize: by weighing the combined variances s_p^2 of the parameters of the formant tracks against the combined χ^2 badness-of-fit, we have obtained a useful smoothness criterion by which we can quantify the smoothness of formant tracks of different lpc analyses.

3. RESULTS

As an example of the improvement that our new algorithm can have, we consider the vowel /o/ as spoken by speaker mpmb0 in sentence si871 of the TIMIT corpus. In the left pane of Fig. 1 the result of

Figure 1: Left panel: first three formants and bandwidths of the vowel /o/ spoken by speaker mpmb0 from TIMIT sentence si871 by a default formant frequency analysis (5 formants, pre-emphasis and 5000 Hz bandwidth). The second formant is displayed in blue colour. Right panel: our algorithm results in an optimal analysis at a bandwidth of 4400 Hz ($F_{min} = 4000$, $F_{max} = 6000$, *nSteps*=21, $\Delta F = 100$).



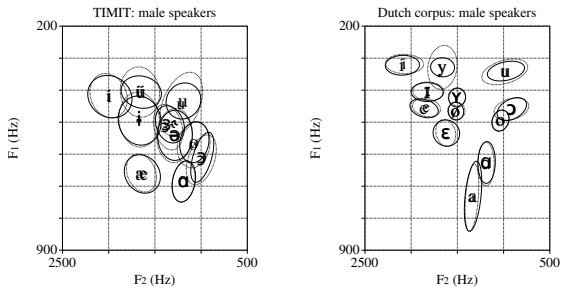
a standard formant frequency analysis from Praat [2] is shown (25 ms analysis window, 2.5 ms time step, pre-emphasis from 50 Hz, burg algorithm and maximum formant frequency of 5000 Hz), where we only have plotted the first three formant tracks as well as their bandwidths. The /o/ sound has a first and second formant that lie close together. However, here the bandwidth of F_1 is large during the first part of this vowel. The second formant is close to the first formant during almost the full duration of the vowel, except in frames 5, 6 and 7 where it suddenly jumps to the position of the F_3. This discontinuity is clearly visible as the values for F_3 also jump upwards. It seems that it would be more natural to interpret the three blue dots in frames 5, 6, and 7 as belonging to F_3 instead of belonging to F_2. But then, how to fill the F_2 gap? There is no definitive answer to that. Another solution exists: the formant analysis outlined above is based on lpc analysis. In lpc analysis we can vary two things: the number of prediction coefficients (the number of formants) and the bandwidth of the signal (the frequency interval where we want our formants to exist). The three formant tracks that are displayed in the right pane of Fig. 1 are the best

tracks that our algorithm could find by varying the bandwidth. This looks like a real improvement.

We will now layout the steps in the algorithm and take this /o/ sound as an example.

1. Decide beforehand on the lowest and the highest bandwidth to analyse (F_{min} and F_{max}). For male voices these frequencies could be 4000 and 6000 Hz. Next decide how many different formant analyses you want to perform (*nSteps*). This fixes the bandwidth frequency step as $\Delta F = (F_{max} - F_{min})/nSteps$.
2. Start: Set the current bandwidth frequency as $F = F_{min}$. Initialize a loop counter ($i = 1$).
3. Downsample the sound to a sampling frequency of $2F$.
4. Perform standard formant analysis with a 25 ms analysis window, a 2.5 ms time step and a pre-emphasis from 50 Hz with the burg algorithm.
5. Model each formant track with a polynomial of low order (in Fig. 1 the chosen order was 3 and only the first three formant tracks were modelled).
6. Calculate the smoothness value W_i according to Eq. (3).
7. Increase the bandwidth: $F = F + \Delta F$ and increase the loop counter: $i = i + 1$
8. If $i < nSteps$ then go to step 3 and continue the analysis with the new bandwidth or else go to End.
9. End: Select the analysis with the smallest W_i value.

Figure 2: The spread of the first and second formant frequencies of labelled vowels in sentences from two data sets and pronounced by male speakers. Left panel: 25909 American-English vowels from TIMIT. Right panel: 4749 vowels from the Dutch corpus.



We have tried our formant frequency algorithm on two acoustic speech corpora, the American English TIMIT data set [5] and a smaller corpus with labelled Dutch vowels [9]. Fig. 2 shows the spread of the first and second formant frequencies, at the mid points of the corresponding vowels, as spoken by male speakers as 1σ ellipses. The ellipses drawn with a dotted line result from the default analysis while the ellipses drawn with a solid line result from our new algorithm. In the right panel the ellipses for the 4749 Dutch vowels, also spoken by male speakers, were drawn with the same convention. The figure shows that for the high vowels /u/ and /ü/ in American-English and also the /u/ and /y/ of Dutch, the ellipses are much smaller for our new analysis algorithm than for the standard analysis with a fixed maximum frequency.

4. CONCLUSION

We have developed a successful formant frequency measurement algorithm that finds the smoothest formant tracks in small vowel-like intervals. The foundation of the algorithm is a successful criterion (Eq. 3) for quantifying the smoothness of formant trajectories.

5. REFERENCES

[1] Adank, P., Van Hout, R., Smits, R. 2004. An acoustic description of the vowels of Northern and Southern Standard Dutch. *J. Acoust. Soc. Am.* 116, 1729–1738.

[2] Boersma, P., Weenink, D. 2015. Praat: doing phonetics by computer [computer program]. Version 5.4.04. Available from http://www.praat.org/.

[3] Escudero, P., Boersma, P., Rauber, A. S., Bion, R. A. H. 2009. A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *J. Acoust. Soc. Am.* 126, 1379–1393.

[4] Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97(5), 3099–3111.

[5] Lamel, L. F., Kassel, R. H., Seneff, S. 1986. Speech database development: Design and analysis of the acoustic-phonetic corpus. *Proc. DARPA Speech Recognition Workshop* 100–109.

[6] Markel, J. D., Gray, A. H., Jr. 1976. *Linear prediction of speech*. Springer Verlag, Berlin.

[7] Nearey, T., Assman, P., Hillenbrand, J. 2002. Evaluation of a strategy for automatic formant tracking. *J. Acoust. Soc. Am.* 112, 2323.

[8] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. 1996. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press second edition.

[9] Van Leussen, J.-W., Williams, D., Escudero, P. 2011. Acoustic properties of Dutch steady-state vowels: Contextual effects and a comparison with previous studies. *Proceedings of the ICPhS XVII, Hong Kong* 1194–1197.

[10] Van Son, R. J. J. H., Pols, L. C. W. 1992. Formant movements of Dutch vowels in a text, read at normal and fast rate. *J. Acoust. Soc. Am.* 92, 121–127.