A new, automated, way to measure vowel formants: Formant detection based on Locally Optimized Spectra*

Ton Wempe
ACLC, University of Amsterdam

Mirjam Ernestus
CLS, Radboud University

Het timbre van een klank wordt meestal weergegeven met zijn formanten. Deze formanten worden bijna altijd gemeten met Linear Predictive Coding (FD_LPC), ook al heeft deze methode nadelen. Dit artikel introduceert een alternatieve methode: formantdetectie gebaseerd op lokaal geoptimaliseerde spectra (FD_LOS). We beschrijven de methode, die geïmplementeerd is als een Praat script. Daarna testen we de methode op basis van 36 constante Nederlandse klinkers door de eerste twee formanten te vergelijken 1) geschat met visuele inspectie van de spectra en berekend met 2) FD_LOS en 3) FD_LPC. Ook testen we de methode door de eerste twee formanten te vergelijken berekend met FD_LOS en FD_LPC voor 475.947 klanken uit drie spraakcorpora, één met informele gesprekken in het Nederlands, één met interviews in het Mandarijn, en één met Engelse zinnen voorgelezen door moedertaalsprekers van het Nederlands. De resultaten laten zien dat de FD_LOS suboptimaal is voor het schatten van de tweede formant. Daarentegen scoort de methode ongeveer net zo goed als FD_LPC voor het schatten van de eerste formant, zonder dat daarvoor parameters ingesteld hoeven te worden. Ten slotte blijkt LOS veel beter dan LPC in staat formanten te detecteren die het resultaat zijn van nasaliteit. De FD_LOS-methode is verder onderzoek waard.

1 Introduction

The quality of a vowel (i.e., its timbre) is typically characterized by the frequencies of its formants, especially its first two formant frequencies. A vowel's formant frequencies are the frequencies of the local maxima of the filter function

* This work tremendously owes to Paul Boersma. Paul spent many, many hours installing Praat on the mini computer on which the second author worked in the 1990s. Moreover, Paul made both of us enthusiastic for detailed phonetic research. In addition, we would like to thank the speakers who produced the vowels that form the basis of our first evaluation of FD_LOS (Section 3).

© Ton Wempe & Mirjam Ernestus

that describes which sound frequencies are amplified as a result of the resonance caused by the shape of the speaker's throat, mouth, and nose (the vocal tract) at the time the vowel is articulated (e.g., Stevens 1998: 31). The frequencies of these maxima can be estimated by hand, for example, by visual inspection of the vowel's Discrete Fourier Transform (DFT) spectrum. In addition, formants can be measured automatically. The, by far, most widely used, automated method is Linear Predictive Coding (LPC, Gray & Markel 1974; for an early overview, see Schroeder 1986), which is also the standard approach in the widely used software package Praat for speech analysis (Boersma & Weenink 2026).

Because of some drawbacks of LPC (Section 2), this article proposes and tests an alternative automated way to estimate formant frequencies (Section 3). This is possible only because of the unparalleled software package Praat (Boersma & Weenink 2026), of which the development was started by Paul Boersma. Together with David Weenink, Paul developed Praat and maintained it on top of his regular work as a teacher and researcher. Praat is of such a high quality because it directly results from Paul's own research on how speech can be analyzed automatically and how phonetic structure can be measured in a robust and theoretically meaningful way, including the analysis of formants (e.g., Escudero et al. 2009), the topic of the present article. Luckily for many researchers, Paul and David were far ahead of their time and made Praat freely available, long before the term Open Access was common in science. In doing so, they have done the scientific community an enormous favor.

We call the new automated way to estimate formant frequencies that we propose in our article Locally Optimized Spectrum (LOS) based formant detection. We first evaluate the merits of this new approach by comparing the values of the formant frequencies of a small number of Dutch vowels that we obtain with LOS based formant detection, by visual inspection of the vowels' spectra, and with LPC (Section 4). We then compare the formant values produced by LPC and LOS based formant detection for speech stretches extracted from three large speech corpora (Section 5). We conclude this article with a discussion of the merits of Locally Optimized Spectrum based formant detection (Section 6).

We implemented LOS based formant detection in a Praat script that can be found at https://www.fon.hum.uva.nl/wempe. In addition, this Praat script, the vowel tokens analyzed in Section 4, the datasets with formant measurements analyzed in Sections 4 and 5, the codes of the scripts that we used to process the data (written for Praat and for the statistical program R, R core team 2026), and the consent forms signed by the speakers of the vowel tokens of Section 4 are all freely available in Wempe & Ernestus (2026).

2 Linear Predictive Coding

LPC was developed almost a century ago and produces an (all-pole) filter that approximates the vocal tract filter function. As such, it serves both speech synthesis and speech analysis, including the estimation of formant frequencies.

However, for estimating formant frequencies, LPC has two important disadvantages. One disadvantage is that the user has to set several parameters for each vowel token that is analyzed, and the best performing parameter settings depend on both the speaker's characteristics and the vowel's quality. The first parameter (the formant ceiling) defines the frequency range that is analysed. A second parameter defines the number of formants that have to be searched for in this frequency range (via the order of the filter, which should be twice the number of poles). These parameters especially can greatly affect the formant values (e.g., Harrison 2013). A third parameter quantifies the speaker's overall spectral slope resulting from the combination of the slopes of the glottal pulse spectrum and the radiation filter at the lips. This parameter avoids that this slope is attributed to the speaker's vocal tract filter. Because the settings of these parameters are so influential, researchers cannot quickly analyze large sets of vowels with LPC: they should determine the LPC parameters for each individual sound separately. In an attempt to start automating this after all, Escudero and colleagues (2009) describe how the formant ceiling of the LPC analysis can be automatically determined based on within-speaker and within-vowel variation.

Another disadvantage of LPC is that, even with correct parameter settings, its output is often not that reliable. This is particularly true for nasal and lateral sounds, because when these sounds are produced, the air exits the vocal tract in more than two ways (in the case of nasals, through both the nose and the mouth, and in the case of lateral sounds, along both sides of the tongue). LPC assumes an all-pole filter without zeroes, while the two parallel sound sources may produce dips (zeros) in the filter function and can therefore only be properly approximated with zeroes in the filter function. The reliability of LPC based formants is further compromised because LPC tends to locate formants closer to strong harmonics than to the centers of the resonances.

Because of these large and systematic errors (Chen, Whalen, & Shadle 2019), Whalen and colleagues (2022) recommend to estimate formants on the basis of the speech signal's spectrogram instead of with LPC. Until now, such an alternative that is also fully automated has not yet been developed.

3 Locally Optimized Spectrum based formant detection (FD_LOS)

LOS based formant detection (FD_LOS) may form an alternative to LPC-based detection (FD_LPC). No parameters have to be set, and researchers can therefore

"blindly" apply the method to large collections of speech data. It estimates formants on the basis of DFT spectra, which are computed for successive short stretches of speech. These spectra are smoothed such that envelope spectra emerge of which the peaks represent formants. The positions of the peaks are determined based on the negative-going zero crossings of the first derivatives of the envelope spectra.

The smoothing of a DFT spectrum is achieved by the convolution of this spectrum with a Gaussian window. The bandwidth of the window depends on the speaker's fundamental frequency (F0) such that it is sufficiently short to avoid averaging over large frequency ranges and sufficiently long to avoid F0 multiples. A value of 1.2/F0 seems optimal.

The durations of the successive stretches of speech for which DFT spectra are computed (i.e., the time windows) should be as short as possible to reflect in detail the changes in the formants but also sufficiently long to attenuate irrelevant changes between successive time windows. A value of five periods seems optimal. If vocal fold vibration is absent, a good value seems to be five periods of the F0 median of the complete speech part for which the formants are tracked. A Hann window is applied to all successive time windows.

Not every peak in the spectral envelope represents a formant. A peak is therefore only interpreted as a formant if the steepness of the leading slope exceeds a threshold value. We found that the optimal threshold varies with the speaker's F0, and an optimal value appears 0.12/F0.

FD_LOS has been formalized in a Praat (Boersma & Weenink 2026) script, which script uses the optimal values mentioned above. The script produces a spectrogram, in which the estimated formant values are indicated with green dots (see Figure 1 for an example). Users can choose not to display the formants detected in the parts of the audio that were produced without vocal fold vibration.

The overlap between the successive windowed speech stretches that are analyzed can be controlled with the parameter "Time resolution of display", which defines the time distance between successive window centers (default value: 5 ms). A value smaller than 5 periods of the F0 produces overlapping time windows. The default (optimal) values for the time window and the spectral window avoid both vertical period lines and horizontal F0 multiples in the spectrogram, as the spectrogram then adapts automatically to the speaker's F0. That is, the spectrogram is then automatically adapted to the speaker's voice, with the exception of pathological voices and creaky voiced parts.

In addition to generating the spectrogram for the speech to be analyzed, the Praat script stores all formant frequencies measured in a Praat "Table" object, which can be saved as a "text file". This table also contains a measure indicating the 'strength' of the formants, defined as the maximum steepness of the upgoing slope of the formant peak in the envelope spectrum. Because the detection of the

formants is just based on the peaks in the envelope spectra (without linguistic knowledge), the formant list also contains peaks in the smoothed spectrum that do not correspond to traditional formants, but indicate, for instance, nasal murmur instead. Formants are numbered in order from low to high.

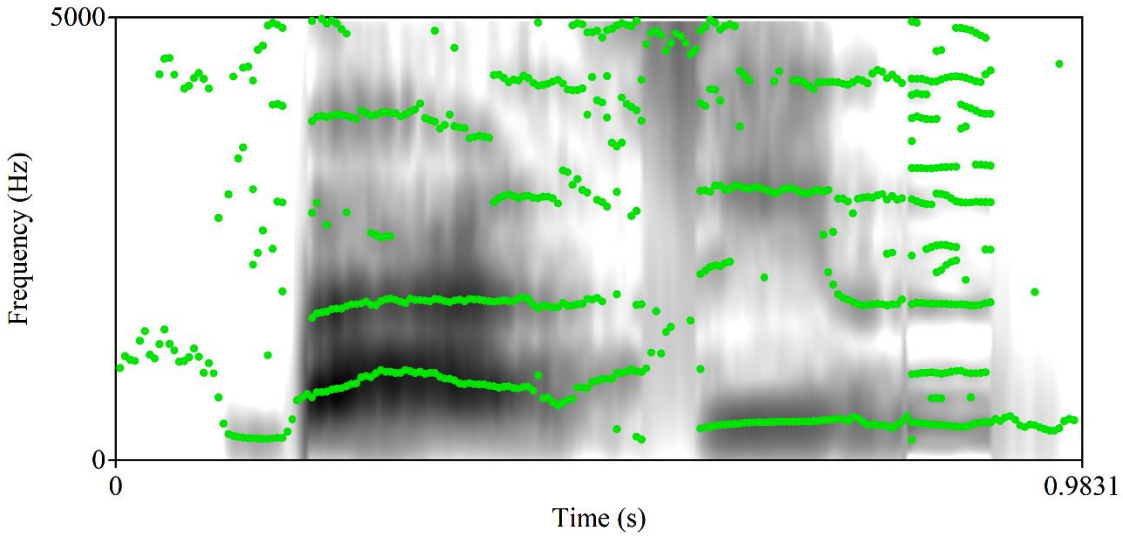


Figure 1: Example of a spectrogram showing the formant values detected by FD_LOS in green. It is a spectrogram of the Dutch word *bazin* /bazɪn/ "female boss".

4 Qualitative and quantitative testing of FD_LOS with Dutch vowel tokens

In order to have a first indication of how well FD_LOS performs, we analyzed the first two formants of 36 invariable tokens of Dutch vowels. We determined the formant values by visual inspection of the vowels' DFT spectra, assuming these values best approximate the actual values. We compared these values with those that we obtained from FD_LOS and FD_LPC as well as with those reported by Adank, van Hout, and Smits (2004). In addition, we directly compared the FD_LOS and FD_LPC values.

4.1. Materials and procedure

We created 36 invariable vowel tokens from natural vowel tokens. Three male and three female native speakers of Dutch, from the Netherlands, each produced six vowels. The speakers include the authors of this article as well as four colleagues who happened to be present at the Erasmus building of the Radboud University on a Friday afternoon, and who were willing to donate their speech without any compensation. The speakers' ages ranged between 25 and 84 years at the time of the recording, and they grew up in the Dutch provinces of Noord-Brabant, Noord-Holland, or Zuid-Holland.

All six speakers produced /i, a, u, ɛ/, while /ɪ, ɔ, ɑ/ were each produced by four speakers. For each natural vowel token, we spliced a single period from the steady state part and copied and concatenated this period multiple times in order to obtain a vowel token of about one second. As the vowel tokens thus created each contain an integer number of periods, do not vary, and are long, their DFT spectra consist of series of clearly distinguishable spectral lines, which facilitates formant detection by visual inspection of the spectra.

Although our analysis focuses on the first two formants, we determined the values of the first three formants (F1, F2, and F3). The reason for this is that FD_LOS or FD_LPC may find additional formants (e.g., a formant below the F1). By reporting the first three formants as detected by these methods, we can see whether they nevertheless correctly identified the F1 and F2 (although then numbered differently). Of course, all measurements were made with the speech analysis software Praat (Boersma & Weenink 2026).

We estimated the formants of each vowel token visually by first consulting the spectrogram of the vowel token to see in which frequency ranges the formants are located (as indicated by dark bands). We then produced the DFT spectrum of the vowel token and estimated the envelope of this line spectrum visually, which approximates the vocal tract filter function for this vowel token. We estimated the frequencies of the maxima in the spectral envelope (and thus of the formants) in the frequency ranges where we had identified the formants in the spectrogram. The maxima were estimated by one author and checked by the other author. The accuracy of the formants estimated visually varied with the speaker's F0. In general, it was difficult to determine where they were exactly located within a range of 20 Hz. We therefore rounded all values to multiples of 10 Hz.

We estimated the formants with FD_LOS with the default settings, see above. For FD_LPC, we could choose from two of Praat's built-in LPC methods: the Burg algorithm (Anderson 1974) and the Robust LPC algorithm (Weenink 2023). The methods produce (nearly) identical formant values for 34 out of our 36 vowel tokens. For the other two tokens, the formant values produced by the Robust method vary much more over time than the values produced by the Burg method, while the vowels do not show any variation over time themselves. We therefore chose to apply the Burg algorithm for all LPC measurements reported in this article. We assumed a pre-emphasis starting at 50 Hz, selected a formant ceiling of 5000 Hz for male speakers and of 5500 Hz for female speakers, and, following the recommendation in the Praat manual, we set the maximum number of formants to five. Although the vowels consisted of identical periods, for some vowels, FD_LPC and FD_LOS showed slight variation in the formant frequencies over time (of about 1%). This variation is the consequence of the fact that successive time windows slightly varied in where they started in a period, and thus in their exact contents. We reported the average values.

Adank and colleagues (2004) reported the average formant frequencies of all Dutch vowels from ten male and ten female speakers from the Netherlands. The vowels /a, i, u/ were produced between an /s/ and a /z/, while /ɛ, ɪ, ɔ, ɑ/ were articulated between two /s/'s. Adank and colleagues obtained their formant measures using the automatic tracking program developed by Nearey, Assmann, and Hillenbrand (2002), which is based on LPC, but allows the user to adjust the formant tracks detected by the tracking algorithm.

4.2 Results

A numeric summary of our results can be found in Table 1. This table shows percentages calculated by dividing the differences between the measurements by the highest value and multiplying the result by 100. The mean differences listed in the table suggest that FD_LPC approaches the visually determined formant values better than FD_LOS.

Table 1: Comparisons of formants values listed by Adank et al. (2004) and those obtained in the present study by visual inspection of the DFT spectra, with FD_LPC, and with FD_LOS.

Source 1	Source 2	**F1 Differences**	**F2 differences**
Adank et al. (2004)	Visual inspection	Mean: 19% Range: 0–54%	Mean: 15% Range: 1–73%
FD_LPC	Visual inspection	Mean: 12% Range: 0-41%	Mean: 13% Range: 3-91%
FD_LOS	Visual inspection	Mean: 19% Range: 0-73%	Mean: 28% Range: 0-80%
FD_LOS	FD_LPC	Mean: 17% Range: 0-72%	Mean: 32% Range: 0-84%

We restrict the quantitative analysis to a minimum because the data set is small. Moreover, while FD_LPC and FD_LOS produced exact formant values for the vowels that we analyzed, the values obtained by visual inspection are mere estimates, and the values reported by Adank et al. (2004) are based on different vowel tokens. We focus on a more qualitative analysis. The cases discussed below are highlighted in the data file and summarized in Table 2.

Table 2: Summary of the more qualitative comparison (see text for an explanation).

Source 1	Compared to	**F1**	**F2**
Adank et al. (2004)	Visual inspection	Much lower (2) Much higher (1)	Much higher (3)
FD_LPC	All other sources	Much higher (1)	Much lower (4)
FD_LOS	All other sources	Missing (2) Nasal murmur (5)	Value between F1-F2 (3) Missing (9)

As shown in Table 1, the formant frequencies estimated visually on the basis of the vowels' spectra more or less correspond to those reported by Adank and al. (2004). Deviations could be expected because we recorded different speakers and these speakers produced the vowels in isolation, unlike Adank and colleagues' speakers. Some of the values differ substantially. For instance, Adank and colleagues reported mean F1s for /ɛ/ and /ɑ/ for male speakers that are (almost) half of the values that we obtained by visual inspection for one male speaker in our dataset. For one other F1 and three F2s, we found the opposite: the values resulting from visual inspection are half of those reported by Adank et al. (F1: one female /ɑ/; F2: a female /i/, a female /ɪ/, and a male /ɛ/).

A few formant frequencies provided by FD_LPC deviate substantially from those estimated visually, those estimated by FD_LOS, as well as from those reported by Adank and colleagues (2004). These are very likely errors. One was a F1 frequency for a male /i/ that FD_LPC estimated to be at least 147 Hz higher than (i.e., at least 1.5 as high as) the values provided by our other sources.

The other clearly deviating formant frequencies computed by FD_LPC are four F2 frequencies that are substantially lower (minimally 500 Hz) than the corresponding F2 frequencies obtained from visual inspection, FD_LOS, and Adank et al. (2004). For instance, FD_LPC found an F2 of 833 Hz for the /ɛ/ of a female speaker, while the other three sources provided F2 frequencies between 1990 Hz and 2465 Hz. In the envelope of this vowel's DFT spectrum, LPC's 833 Hz formant is located at the slowly decreasing trailing edge of the maximum representing the F1. Our other sources did not interpret the form of this trailing edge as sufficient evidence for a formant. The three other LPC F2 values that were much lower than we obtained from our other sources are from another female speaker's /i, ɪ, and ɛ/.

These four F2 errors form over 10% of the F2 values computed by FD_LPC. This percentage is very likely a major underestimation of the real percentage of errors because we left out of consideration F2 values computed by FD_LPC that substantially deviated from the values estimated visually and by FD_LOS, but that corresponded more or less with Adank et al. (2004)'s estimates.

FD_LOS produced more F1 errors than LPC. First, it missed two F1s (for a female /a/ and a male /ɔ/), reporting values that are too high and that better approximate the corresponding F2s estimated by visual inspection. Second, it detected five F1s that are substantially lower than the values we obtained from our other sources (two male /a/s, a male /ɛ/, a male /ɑ/, and a female /ɪ/). These five F1 values range between 247 Hz and 297 Hz and the corresponding maxima are visible in the spectrograms as light formant bands. They probably reflect a (slight) nasal articulation of the vowel tokens. As expected, FD_LOS is much better than FD_LPC in detecting nasality.

For two out of the five vowel tokens for which FD_LOS identified nasal murmur (for a male speaker's /ɛ/ and /a/), FD_LOS assigned values to the F2s that well correspond to the F1 values obtained with visual inspection and with FD_LPC. For these vowel tokens, FD_LOS found nasal murmur in addition to the regular formants, while for the other three tokens, FD_LOS seems to have replaced the F1 by nasal murmur.

FD_LOS not only produced erroneous F2 values when nasal murmur was present. It detected an additional formant between the F1 and F2 for three vowel tokens (a male /i/, a female /i/, and a female /ɛ/): the F3 values computed by FD_LOS for these vowel tokens correspond to the F2 values determined by Adank et al. (2004), by FD_LPC, and with visual inspection. In addition, FD_LOS missed the F2s for nine vowel tokens: It produced F2 values for these tokens that are at least 500 Hz higher than those determined by Adank et al. (2004), by FD_LPC, and with visual inspection, and that better fit these sources' F3 values. FD_LOS only detects formants that are signaled by peaks in the envelope of the DFT spectrum. It misses F2s that are only visible in the spectral envelope in that they make the slope of the trailing edge of the F1 less steep at that position, instead of being local maxima.

In conclusion, nasal murmur was better detected by FD_LOS than by FD_LPC. This capability of FD_LOS explains at least partly why the differences between the F1 values obtained by visual inspection of the spectra differ more from those computed with FD_LOS than with FD_LPC. If FD_LOS detects nasal murmur, either its F2 value should be interpreted as its F1 value (and its F3 value as its F2 value, and so on), or it misses the F1. FD_LOS appears worse than FD_LPC in detecting F2s because it can only identify formants if they form separate peaks in the spectra. FD_LPC, however, also produced F2 errors for at least 10% of the vowel tokens.

5 Comparing LPC and LOS based formants from large speech corpora

The comparison of FD_LPC and FD_LOS in Section 4 is based on vowel tokens that were carefully articulated in isolation. We continue evaluating FD_LOS by

comparing its results with those from FD_LPC on the basis of three large speech corpora, each representing a different language and degree of formality.

5.1 Materials and procedure

We analyzed the speech in three corpora. The first corpus is the Ernestus Corpus of Spontaneous Dutch (Ernestus 2000), consisting of ten informal conversations, each 90 minutes long, between male speakers, who were friends or direct colleagues. The second corpus is Fon (2004)'s Taiwan Mandarin speech corpus, consisting of eight hours of answers from eight male and eight female speakers to interview questions. The third corpus consists of the speech from 96 female native speakers of Dutch with a B2 level in English (Council of Europe 2021), carefully articulating 96 sentences in English (Troncoso Ruiz 2022, Chapter 3).

We extracted the speech files from the corpora by means of Praat scripts. We only analyzed those speech stretches that, according to the corresponding TextGrids with orthographic transcriptions, contain speech. We excluded small parts of the Dutch corpus (Speaker pair 1_2: Parts 1, 3, 10; Speaker pair pp8_X2: Part 2; Speaker pair X5_X6: Part 2; Speaker pair 15_16: Part 14) because they do not contain well detectable speech or because the processing of the TextGrids resulted in errors. From the Mandarin corpus, we excluded LJS_GY, XMC_GY and ZWH_GY, because each of these resulted in over 100,000 speech stretches, which Praat cannot contain at the same time. Because not all three corpora come with transcriptions at the sound level, we tried to localize the vowels in the speech stretches by applying the Praat script by de Jong & Wempe (2009). This script identifies the maxima of intensity envelopes, which are mostly syllable nuclei.

We determined the F1, F2, and F3 with FD_LPC and FD_LOS as described in Section 4.1. The resulting dataset contains the formant measurements for the 642,304 speech stretches identified as syllable nuclei by the Praat script (de Jong & Wempe 2009) and identifies the language and the speaker's gender. In addition, it provides the bandwidth for each formant detected by FD_LPC and the strength for each formant detected by FD_LOS. We focused our analysis on the formants estimated for the 483,170 voiced stretches of speech.

5.2 Results

Unlike FD_LPC, FD_LOS did not detect formants for all tokens in the dataset: it failed to find F1 values for 1.5% of the tokens and F2 values for 6% of the tokens. FD_LOS only reports formant values when the DFT spectrum shows peaks. F2 values were most often absent for the Dutch spontaneous speech corpus (14,613 tokens), followed by the Mandarin interview corpus (12,403 tokens), followed by the English read aloud corpus (2018 tokens).These differences between the corpora suggest that FD_LOS only produces formant values for stretches of

speech with (more or less) clear vowels. We discarded the tokens for which FD_LOS did not produce F1 and F2 values. Our final dataset consists of 475,947 tokens: 116,410 from the Dutch corpus, 169,827 from the English corpus, and 189,710 from the Mandarin corpus.

Table 3 presents a numeric summary of the differences in the F1 and F2 values generated by FD_LOS and FD_LPC. Table 4 specifies the inaccurate formant values.

Table 3: Comparison of formant values generated by FD_LOS and FD_LPC for the dataset based on three speech corpora.

	F1	**F2**
Mean difference	96.3 Hz	507.7 Hz
Median difference	14.0 Hz	82.0 Hz
% of tokens with difference >10% (of the highest of the two values)	19.2 %	30.4 %

Table 4: Reasons for inaccurate F1 and F2 values in the dataset based on the three corpora.

	F1	**F2**
% of (additional) tokens identified by FD_LOS as having nasal murmur	2.6%	
% of improbably high values produced by FD_LOS	1.5%	1.5%
% of LPC formant values (F) with bandwidths > (F/2)	7.4%	14.5%

As Table 3 shows, the F1 values produced by the two measurement methods are similar. The mean difference in F1, however, varies greatly depending on the corpus: 167.4 Hz, 44.3 Hz, and 99.3 Hz, for the Dutch, English, and Mandarin corpora, respectively.

One of the reasons why FD_LOS and FD_LPC report different F1 values for some stretches of speech is, as also documented in Section 4, that the two methods greatly differ in their ability to detect nasal murmur. FD_LOS produced more than twice as many F1s with values below 300 Hz than FD_LPC did (for 53,967 versus 23,898 tokens). For 2.6 % of all F1s, FD_LOS found values that are both lower than 300 Hz and at least 200 Hz lower than the corresponding F1 values obtained with FD_LPC (see Table 4).

Further, the two formant detection methods differ in the number of F1 values above 2000 Hz that they detect. Such high values are improbable, especially when they are not supported by alternative formant detection methods. They indicate that the real F1s were missed. FD_LPC produced 34 F1s that are both above 2000 Hz and minimally 500 Hz above the FD_LOS values. In contrast, FD_LOS produced 7,237 F1 values that are both above 2000 Hz and at least 500 Hz above the values produced by FD_LPC (1.5% of the F1 values; see Table 4).

The average of these FD_LOS values is 3,055 Hz. Interestingly, only 9.6% of these exceptionally high F1 values occur in the English corpus.

For F2, Table 3 shows that the difference between the values produced by FD_LOS and FD_LPC is much greater than for the F1 (see Table 3). As for the F1, the mean F2 difference varies greatly depending on the corpus: 700.0 Hz, 355.4 Hz, and 526.0 Hz, for the Dutch, English, and Mandarin corpora, respectively.

F2 values above 4000 Hz are likely to be errors. FD_LPC produced two of these. In contrast, FD_LOS produced 7,247 of such values (1.5%, see Table 4): 3,816 for the Dutch, 2,834 for the Mandarin, and 597 for the English dataset.

It is an open question whether large differences between FD_LOS and FD_LPC are due to errors made by FD_LOS or by FD_LPC. It may well be that FD_LPC produced a high number of errors, for reasons mentioned in Section 2 and as shown in Section 4.2. In addition, we found that the bandwidth for 7.4% of the F1s and 14.5% of the F2s computed by FD_LPC are at least half of the formant values themselves (see Table 4). This indicates that those formants represent rather flat maxima in the filter function of the vocal tract, and may therefore not be proper F1 and F2 values.

6 General discussion and conclusion

This article presents a new method for measuring formants that is based on estimating the positions of the maxima in the envelope of the DFT spectrum. It is implemented in a Praat script such that the default values of the parameters are sensitive to the speaker's local fundamental frequency, which produces the best results, as shown by informal piloting of the method. Hence the method's name: Locally Optimized Spectra (LOS) based formant detection (FD). Because no speaker-specific or vowel token specific settings of parameters are necessary, the method (FD_LOS) can be applied "blindly" to large speech datasets, which is an advantage above the commonly used LPC method.

We presented two tests of the new method. First, we tested it against invariable Dutch vowels, for which we determined the first and second formant values by visual inspection of the DFT spectrum, and by comparing these values with those that we obtained from FD_LOS and LPC (FD_LPC) as well as with those reported by Adank and colleagues (2004). Second, we compared the outputs of FD_LOS and FD_LPC for over 475,947 stretches of speech from three existing speech corpora representing different languages and degrees of formality.

A first result is that FD_LOS largely outperforms FD_LPC in detecting nasal murmur: while FD_LPC ignores nasal murmur, FD_LOS reports it as the F1. FD_LOS detected nasal murmur for five vowels in the invariable dataset (Table 2) and it detected almost twice as many vowels with low formants (below

300 Hz) than LPC did in the dataset of three speech corpora (in 2.6% of the vowel tokens, Table 4). This makes FD_LOS a good method for studying nasality.

Both FD_LPC and FD_LOS produced more errors for the F2 than for the F1. For over ten percent of the invariable vowels, FD_LPC produced F2s that greatly deviate from the values form the other three sources (visual inspection, FD_LOS, and Adank et al. 2004, see Table 2). The actual FD_LPC error rate is probably higher because we ignored F2 values that correspond to the values reported by Adank et al. but that differ greatly from those we obtained by visual inspection and with FD_LOS. Furthermore, the data from the three speech corpora show that LPC produced 14.5% of the F2 values with very large bandwidths (see Table 4; ignoring the tokens for which FD_LOS produced no formant values at all), which also indicates that many F2 values are not accurate.

Nevertheless, FD_LPC appears to outperform LPC_LOS in estimating F2 frequencies. The values obtained by visual inspection showed a higher average difference with the FD_LOS than with the FD_LPC values (Table 1). Further, for the three speech corpora, FD_LOS produced impossible values of over 4000 Hz for 1.5% of the F2s (Table 4), while FD_LPC did so for only two F2s.

There are at least two reasons why FD_LOS performs worse than FD_LPC detecting F2s. First, if FD_LOS detects nasal murmur, it is labeled as the F1, and the F2 value should then be interpreted as the F1, and the F3 value as the F2. A second likely reason for the underperformance of FD_LOS is that FD_LOS is designed such that it can only detect formants that form well discernible maxima in the vowel's spectral envelope, while some F2s seem to manifest themselves as flattish maxima due to incorporation by the F1s, or as locally low maxima. Future research should show whether, and, if so, how, FD_LOS can be adapted to recognize these manifestations of the F2 as well.

FD_LOS' performance for the F2 appears to be substantially dependent on the speech corpus: it performs rather well for the English sentences read aloud by non-native speakers, less so for the interviews in Mandarin, and worst for the informal conversations between acquaintances in Dutch. Possibly, this has something to do with how we tried to identify the vowels in our corpora: our method may have more often identified other sounds than vowels (e.g., voiced consonants or non-speech sounds) as syllable nuclei, the more informal the speech was. FD_LOS may have more difficulties determining the formants for such sounds as they are more likely to have a flat spectrum. We leave the question of whether this is true and, if so, why, for future research.

In conclusion, FD_LPC outperforms the current version of FD_LOS in accuracy of measuring F2s. In contrast, the current version of FD_LOS outperforms FD_LPC in two other respects: it much better detects nasal murmur, and it performs similarly to FD_LPC in estimating F1 values but without parameters whose values have to be determined on the basis of the characteristics

of the speaker (and the vowel token). This makes further research on FD_LOS promising.

7 References

Adank, P., Van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America 116*(3): 1729–1738. https://doi.org/10.1121/1.1779271

Andersen, N. (1974). On the calculation of filter coefficients for maximum entropy spectral analysis. *Geophysics 39*(1): 69–72. https://doi.org/10.1190/1.1440413

Boersma, P. & Weenink, D. (2026). *Praat: doing phonetics by computer* [Computer program]. Version 6.4.61. https://praat.org/

Chen, W. R., Whalen, D. H., & Shadle, C. H. (2019). F0-induced formant measurement errors result in biased variabilities. *The Journal of the Acoustical Society of America 145*(5): EL360–EL366. https://pubs.aip.org/asa/jasa/article/145/5/EL360/671477

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://www.coe.int/en/web/common-european-framework-reference-languages/

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetic interface*. Holland Institute of Generative Linguistics, Utrecht. https://repository.ubn.ru.nl/bitstream/handle/2066/75101/304805.pdf

Escudero, P., Boersma, P., Rauber, A. S. & Bion, R. A. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America 126*(3): 1379–1393.

Fon, J. (2004). *A preliminary construction of Taiwan Southern Min spontaneous speech corpus*. Technical report, NSC-92-2411-H-003-050, National Science Council, Taiwan.

Gray, A. & Markel, J. (1971). A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing 22*(3): 207–217. https://ieeexplore.ieee.org/document/1162572

Harrison, P. (2013). *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. [Unpublished doctoral dissertation]. University of New York. https://etheses.whiterose.ac.uk/id/eprint/7393/1/Harrison%20-%20PhD%20Thesis.pdf

Jong, de, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41(2): 385–390. https://link.springer.com/article/10.3758/BRM.41.2.385

Nearey, T. M., Assmann, P. F., & Hillenbrand, J. M. (2002). Evaluation of a strategy for automatic formant tracking. *The Journal of the Acoustical Society of America 112*(5_Supplement): 2323–2323. https://doi.org/10.1121/1.4779372

R core team (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Schroeder, M. (1986). Predictive coding of speech: Historical review and directions for future research. *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing 11*, 3157–3164. https://ieeexplore.ieee.org/abstract/doument/1168600

Stevens, K. N. (2000). *Acoustic phonetics*. MIT Press.

Troncoso Ruiz, M. A. (2022). *Non-native phonetic accommodation in interactions with humans and with computers.* Radboud University Nijmegen: MPI. https://repository.ubn.ru.nl/bitstream/handle/2066/250489/250489.pdf?sequence=1&isAllowed=y

Whalen, D. H., Chen, W. R., Shadle, C. H., & Fulop, S. A. (2022). Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986). *The Journal of the Acoustical Society of America 152*(2): 933–941. https://doi.org/10.1121/10.0013410

Weenink, D. (2023). Improved formant frequency measurements in Praat. *Proceedings of the ICPhS* 20: 704-707. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/463.pdf

Wempe, T., & Ernestus, M.T.C. (2026). Formant detection based on Locally Optimized Spectra. Version 1. Radboud University. [dataset]. https://doi.org/10.34973/pzdj-js20

Ton Wempe: a.g.wempe2@contact.uva.nl

Mirjam Ernestus: Mirjam.Ernestus@ru.nl