

# **Pitch-Synchronous Analysis of Speech**

Author:

**C. Julian Chen**

Department of Applied Physics and Applied Mathematics

Columbia University

New York, NY 10027

Presented at the 2019 Annual Symposium of The Voice Foundation

June 1, 2019, The Westin, Philadelphia, PA

# Pitch-Synchronous Analysis of Speech

## Abstract

**Objective:** To develop an accurate and reliable method for spectroscopic analysis of speech for acoustic phonetics and its applications in speech technology.

**Methods/Design:** The method of pitch-synchronous spectroscopic analysis of speech is based on the definition of formants in the original papers of Ludimar Hermann and the transient theory of speech production presented by Peter Ladefoged in *Elements of Acoustic Phonetics*. First, the pitch marks, i.e. the instants of starting pulses in pitch periods, are extracted from speech signals using an asymmetric window. The speech signal is then segmented into pitch periods. After the two ends of the segment of signals in each pitch period are equalized, a Fourier analysis is applied to obtain the timbre spectrum of each pitch period. To reduce the effect of noise and to improve accuracy, a running average of the timbre spectra over a number of consecutive pitch periods is executed. The resulting timbre spectrum can be displayed as a graph or the numerical records can be output for further analysis.

**Results:** To demonstrate the universal applicability of the concept and the method, a user-friendly software of pitch-synchronous spectroscopy of speech is developed, which can display the pitch-synchronous spectrogram and the timbre spectrum in each pitch period individually. For academic users, an evaluation version can be acquired from Columbia Technology Ventures for free<sup>12</sup>. The package contains a number of examples of various languages selected from the UCLA Phonological Segment Inventory Database to demonstrate that the software can be used to analyze sounds of the world's languages under various recording conditions.

**Conclusions:** By first segmenting speech signals into natural units -- pitch periods, then making Fourier analysis, the richness and accuracy of spectroscopic information are substantially better than the traditional pitch-asynchronous formant tracking using linear predictive coding method. The richness of the pitch-synchronous spectroscopy of speech suggests a more detailed study of the voice production mechanism beyond a small number of formants, by comparing with the images of vocal tracts from real-time magnetic resonance imaging using numerical integration of the time-dependent wave equation inside the vocal tract.

Keywords: Speech, Acoustic phonetics, Spectroscopy, Formants, Pitch period, Graphical display.

# Pitch-Synchronous Analysis of Speech

## Abstract

**Objective:** To develop an accurate and reliable method for spectroscopic analysis of speech for acoustic phonetics and its applications in speech technology.

**Methods/Design:** The method of pitch-synchronous spectroscopic analysis of speech is based on the definition of formants in the original papers of Ludimar Hermann and the transient theory of speech production presented by Peter Ladefoged in *Elements of Acoustic Phonetics*. First, the pitch marks, i.e. the instants of starting pulses in pitch periods, are extracted from speech signals using an asymmetric window. The speech signal is then segmented into pitch periods. After the two ends of the segment of signals in each pitch period are equalized, a Fourier analysis is applied to obtain the timbre spectrum of each pitch period. To reduce the effect of noise and to improve accuracy, a running average of the timbre spectra over a number of consecutive pitch periods is executed. The resulting timbre spectrum can be displayed as a graph or the numerical records can be output for further analysis.

**Results:** To demonstrate the universal applicability of the concept and the method, a user-friendly software of pitch-synchronous spectroscopy of speech is developed, which can display the pitch-synchronous spectrogram and the timbre spectrum in each pitch period individually. For academic users, an evaluation version can be acquired from Columbia Technology Ventures for free<sup>12</sup>. The package contains a number of examples of various languages selected from the UCLA Phonological Segment Inventory Database to demonstrate that the software can be used to analyze sounds of the world's languages under various recording conditions.

**Conclusions:** By first segmenting speech signals into natural units -- pitch periods, then making Fourier analysis, the richness and accuracy of spectroscopic information are substantially better than the traditional pitch-asynchronous formant tracking using linear predictive coding method. The richness of the pitch-synchronous spectroscopy of speech suggests a more detailed study of the voice production mechanism beyond a small number of formants, by comparing with the images of vocal tracts from real-time magnetic resonance imaging using numerical integration of the time-dependent wave equation inside the vocal tract.

## Introduction

Spectroscopic analysis of speech is important in many fields of research and technology. In acoustic phonetics, the individual sounds of speech of the world is compared through the spectra in frequency domain.<sup>1,2</sup> In speech recognition, the spectroscopy of incoming speech is compared with a database of small spectral units with possible phonemes.<sup>3,4</sup> In speaker identification and speaker verification, the sound of the incoming speech is compared with a database of many

persons.<sup>5,6</sup> Especially for forensic phonetics, the depth and scope of information, the accuracy and objectiveness of the measurements, are of paramount importance. Some of the existing technologies are still based on the legacy telephone signals with 8 bit depth and 8 kHz sample rate. Much advance can be achieved by using modern audio acquisition and transmission technology coupled with improved computing power. Here we introduce pitch-synchronous spectroscopy of speech, which is much more accurate and reliable than the existing pitch-asynchronous spectroscopic analysis such as linear predictive coding.

The organization of this article is as follows. First, we present the definition of formants following the papers of Ludimar Hermann.<sup>7,8</sup> Next, we discuss the nature of the sound of speech following Peter Ladefoged's *Elements of Acoustic Phonetics*.<sup>9</sup> Both narratives point to the importance of pitch-synchronous analysis of speech. Then, practical methods and procedures to segment speech into pitch periods and making spectroscopic analysis are presented. Noise cancelation using running average over consecutive pitch periods is introduced. Finally, a user-friendly speech analyzer is presented, which is free for academic users. In Appendix II, we discuss the pros and cons of formant tracking using linear predictive coding.<sup>10-12</sup>

## Hermann: The Definition of Formants

Quantitative study of the spectroscopy of speech started in late nineteenth century after Edison invented phonograph in 1877. Waveforms of speech can be recorded as mechanical grooves on wax or soft-metal films. In late nineteenth century, German physiologist Ludimar Hermann further amplified the mechanical grooves and fixed it on photographic papers, then performed a systematic study of the waveforms using Fourier analysis. From 1889 to 1911, he published 12 original research papers and one review paper in *Pflüger Archiv*, now *European Journal of Physiology*. Especially, in an 1894 paper entitled *Phonographische Untersuchungen VI. Nachtrag zur Untersuchung der Vocalcurven (Phono-photographical Studies VI. Supplements on the Research of Vowel Waveforms)*,<sup>9</sup> details of the procedures are disclosed, and a term Formant is defined. An annotated version of that paper is attached as an appendix.

Hermann's method of spectroscopic analysis of speech was as follows. The speech sounds were first recorded using an Edison phonograph. The grooves were optically amplified and fixed on photographic papers. Using his apparatus, the sizes of the curves were enlarged to more than 10 cm. The typical dimension of the curves was about 5 cm. Using a glass with millimeter grids, the curves were manually digitized into numerical tables. The accuracy of such manual reading is about a quarter millimeter. Because a typical pitch period is 5-10 msec, with 5 cm length, the time resolution is about 0.1 msec. In modern technical jargon, the sound wave was quantized to 8 bit depth at a sample rate of about 8 kHz. Examples of the amplified waveforms are shown in Figure 1. As shown, Each pitch period starts with a strong pulse, then decays in the pitch period. The waveforms of neighboring pitch periods are similar. The waveforms are segmented into pitch periods. Denoting the pitch period as  $T$ , the pitch frequency is defined by Equation (1):



FIGURE 1. Examples of speech waveforms studied by Hermann, adapted from Figure 31 of Scripture's *The Elements of Experimental Phonetics*,<sup>8</sup> on page 40.

$$f_o = \frac{1}{T}. \quad (1)$$

Hermann then performed Fourier analysis on the waveforms of various vowels. Details were described in the Appendix of that 1894 paper.<sup>7</sup> Assuming that each pitch period has  $N$  points in time, the speech signal  $s(n)$  can be represented by Fourier series, see Equation (2),

$$s(n) = \frac{a_0}{2} + \sum_{r=1}^{r \leq (N/2)} \left[ a_r \cos \frac{2\pi rn}{N} + b_r \sin \frac{2\pi rn}{N} \right], \quad (2)$$

here the coefficients  $a_r$  and  $b_r$  are computed from the waveforms of *a single pitch period*,

$$\begin{aligned} a_r &= \frac{2}{N} \sum_{n=1}^{n < N} s(n) \cos \frac{\pi nr}{N}, \\ b_r &= \frac{2}{N} \sum_{n=1}^{n < N} s(n) \sin \frac{\pi nr}{N}. \end{aligned} \quad (3)$$

The *amplitude spectrum* of the sound is defined as

$$c_r = \sqrt{a_r^2 + b_r^2}. \quad (4)$$

Initially, those amplitude spectral coefficients are calculated on the integer multiples of pitch frequency, defined by Equation (1). If at order number  $r$ , the amplitude reaches maximum, then a characteristic peak frequency of the vowel is calculated by

$$F = rf_o. \quad (5)$$

Hermann further assumed that the amplitude spectrum exists for all values of frequency, which should be interpolated from the values at the integer multiples of the pitch frequency. Therefore, in general, the order number  $r$  (Ordnungszahl in German) in Equation (5) is a fractional number. Through a study of large number of cases, Hermann found that those peak frequencies in the

spectra are characteristic to the vowel and independent of pitch frequency. Those characteristic frequencies was defined by Hermann as the *formants* of that vowel.

For some vowels, such as [a], [o], and [u], a single formant is sufficient to indentify its nature. For some vowels, more than one formants are observed, such as [ae]. On page 269, Hermann listed four cases of vowel [ae] on pitch frequencies 147 Hz, 165 Hz, 196 Hz and 261 Hz. To a standard variation of 10%, he found two formants at average frequencies 609 Hz and 1548 Hz, independent of the pitch frequency. Those values of the formants of [ae] agrees with modern measurements, for example, 660 Hz and 1720 Hz, on page 45 of Robiner and Schafer.<sup>3</sup>

In nineteenth century, the Fourier coefficients in Equation (3) were computed manually. It was an enormous endeavor. Details of the numerical calculation are described in Appendix I, pages 561-574 of E. W. Scripture's *The Elements of Experimental Phonetics*.<sup>8</sup> On modern computers, those type of computations takes only a fraction of a microsecond. Moreover, the accuracy of modern sound recording, typically with 16 bit resolution at a sample frequency of 44.1 kHz, is far better than the mechanical grooves of the Edison phonograph. However, Hermann's scientific procedure and the definition of formants are still valid. Here is a brief summary:

- (1) Cut out the waveform of a single pitch period from the sound recording. Obviously, the starting value and the ending value should be approximately equal to ensure continuity.
- (2) Compute the Fourier coefficients from the waveform of a single pitch period, in both sine and cosine, according to Equation (3).
- (3) Compute the amplitude spectrum by combining the sine and cosine components according to Equation (4), to find the amplitude spectrum at the integer multiples of the pitch frequency  $f_0$ .
- (4) Interpolate the discrete spectrum to a continuous scale of frequency.
- (5) Find the peaks in the continuous curve of amplitude versus frequency.

As a historical fact, Ludimar Hermann pioneered *pitch-synchronous analysis of speech*.

## **Ladefoged: The Sound of Speech**

In Chapter 7 of Ladefoged's *Elements of Acoustic Phonetics*,<sup>9</sup> entitled *The Production of Speech*, the waveforms of his own voice are analyzed to show the nature of the sound of speech. First, his sound of vowel [ɔ] is analyzed. It consists of a series of damped waves, each starts with a pulse from the vocal folds, see Figure 2(A). The frequency of each damped wave, Figure 2(A)(1), is about 500 Hz. The waveform can be approximated by Equation (6)

$$s(t) = Ce^{-\beta t} \sin 2\pi Ft, \tag{6}$$

where  $F=500$  Hz is the formant frequency,  $\beta$  is the decay constant, and  $C$  is a constant indicating its intensity. The vocal folds emit a series of pulses to trigger those damped waves. In Figure 1(A), (2) through (4), three cases are displayed: with pulse rate 100, 120 or 150 times a second. The pitch frequency of sound [ɔ] is 100 Hz, 120 Hz or 150 Hz, respectively.

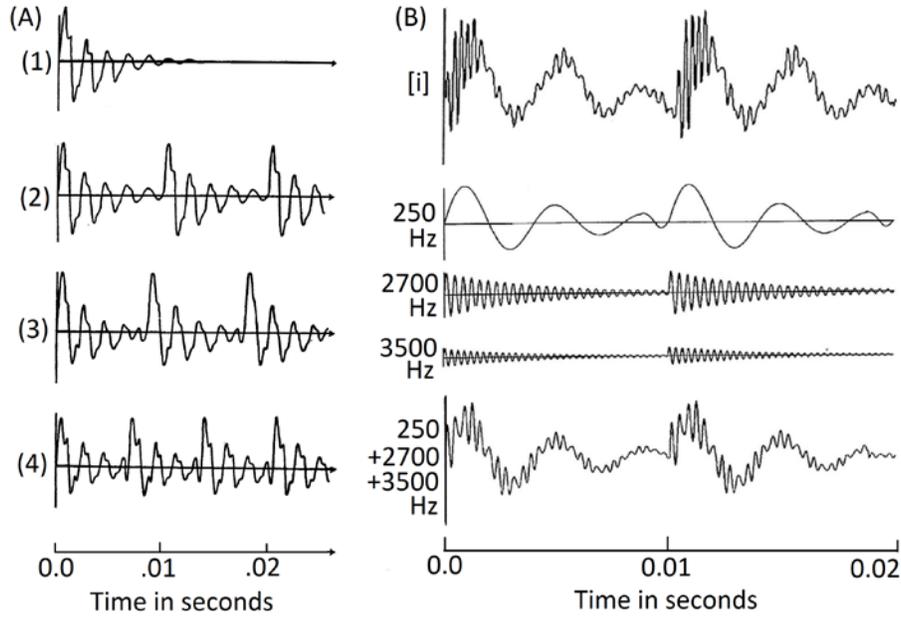


FIGURE 2. Nature of speech sound as explained by Ladefoged in Chapter 7 of *Elements of Acoustic Phonetics*<sup>9</sup>

Next, Ladefoged analyzed his sound of vowel [i], see Figure 2(B). Each pulse from the vocal folds triggers three damped waves, in 0.25 kHz, 2.7 kHz, and 3.5 kHz, respectively. Here the vocal folds emit 100 pulses per second, producing a vowel sound [i] of pitch frequency 100 Hz. Each elementary wave can be represented by Equation (7),

$$s(t) = \sum_{n=1}^3 C_n e^{-\beta_n t} \sin 2\pi F_n t. \quad (7)$$

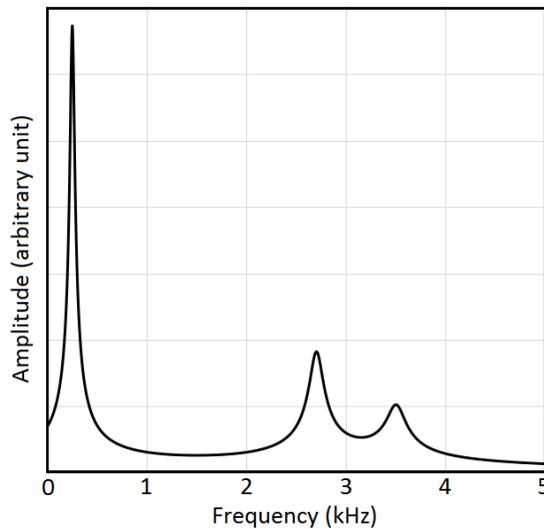


FIGURE 3. Amplitude spectrum of vowel [i] described by Ladefoged. To fit the curves in Figure 2(B), the parameters are:  $C_1=1.0$ ,  $C_2=0.75$ ,  $C_3=0.5$ ;  $\beta_1=0.15$ ,  $\beta_2=0.45$ ,  $\beta_3=0.6 \text{ msec}^{-1}$ .

Here we study the spectrum of vowel [i]. Assuming that the three elementary damped waves are approximately independent, a Fourier transform produces an amplitude spectrum as

$$A(f) = \sum_{n=1}^3 \frac{C_n}{\sqrt{\beta_n^2 + 4\pi^2(f - F_n)^2}}. \quad (8)$$

A derivation of Equation (8) is in Appendix I. A graphical representation is shown in Figure 3. As shown, the amplitude spectrum of vowel [i] peaks at three characteristic frequencies, or formants, of 0.25 kHz, 2.7 kHz, and 3.5 kHz.

The description of the voice by a group of damped waves are fairly accurate and universal. Figure 4 shows the waveform and spectrum of a recent recordings, a vowel [i] spoken by a male specialist, extracted from database King-TTS-012 of SpeechOcean.<sup>10</sup> Waveforms of individual pitch periods are manually cut out, then a Fourier analysis is performed. As shown, even as the pitch frequency varies from 263 Hz to almost zero Hz, where the pitch period becomes very long, the timbre spectrum is virtually unchanged. In other words, the Fourier transform of the signal in a short pitch period (less than 4 msec) is virtually identical to the Fourier transform of a damped wave by letting it decays freely to zero. The spectrum in Figure 4 shows four formants, one of a lower frequency, and three of frequencies three octaves higher. A survey of many samples of vowel [i] shows a similar pattern: a strong formant at a fairly low frequency, plus a group of high-frequency formants, typically three, but sometimes two, four, or five. The details of the high-frequency group is important for speaker verification and forensic acoustics.

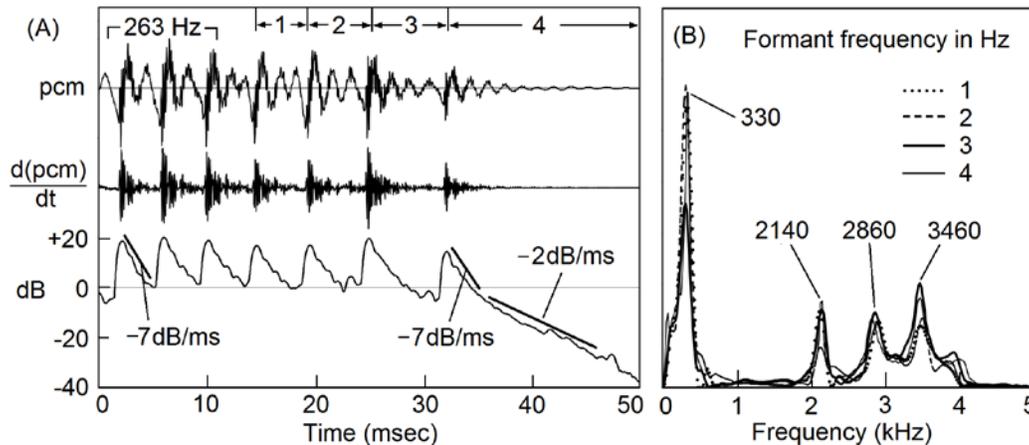


FIGURE 4. Waveform and spectrum of vowel [i] from a standard speech database: SpeechOcean King-TTS-012, sentence 004419, 1.938 seconds to 1.988 seconds.

The observed independence of the length of the pitch period can be explained by Figure 5. The waveform in a single pitch period, in (E), is the sum of the current damped wave, plus the tails of the damped wave from the previous pitch periods. Therefore, the Fourier transform of the signal in a single pitch period is identical to the Fourier transform of a complete damped wave.

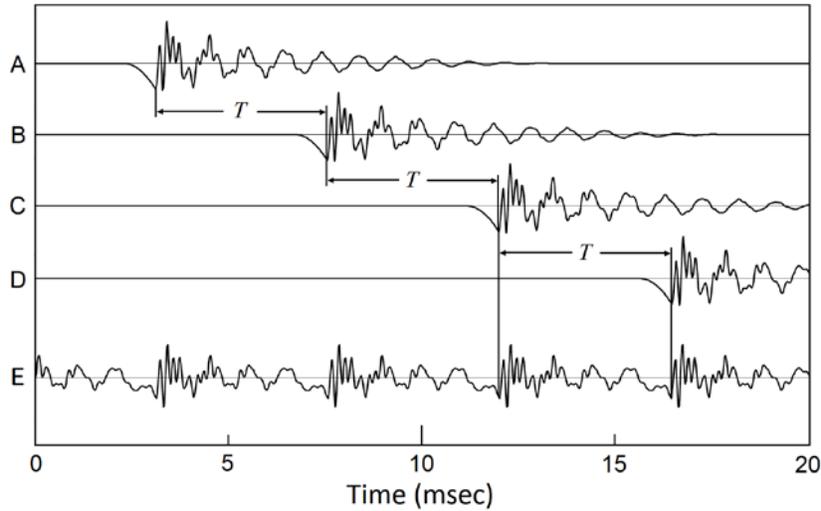


FIGURE 5. The waveform of a single pitch period as the sum of the current damped wave and the tails of the damped waves of previous pitch periods.

The nature of the speech given by Ladefoged in Equation (7) and the spectrum in Equation (8) can be considered as a definition of formants from a speech production point of view.

## Automatic Pitch-Period Segmentation

Ludimar Hermann started his spectroscopic analysis of speech sound by segmenting waveforms into pitch periods by hand. For the case of a few pitch periods, such as displayed in Figure 4, it is still doable. To process large quantity of speech signals, automate pitch-period segmentation is a necessity. The nature of speech signals described by Ladefoged, cited in the previous section, is the basis of automatic pitch-synchronous segmentation of speech signals.

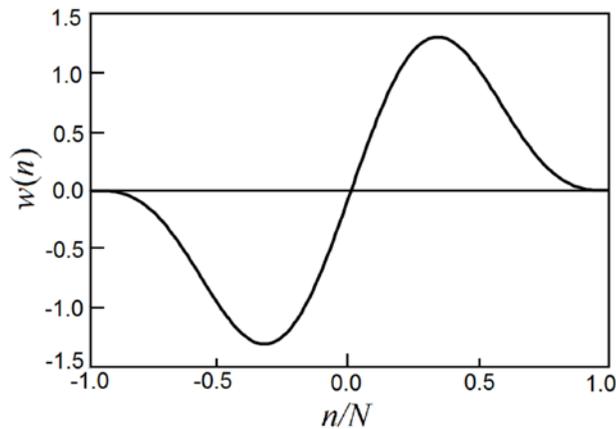


FIGURE 6. An asymmetric window function for the generation of excitation profiles.

As shown in Figures 2 and 4, the signal of each pitch period starts with a strong pulse, then decays, until the next strong pulse comes up. Therefore, the starting point of each pitch period is characterized by a surge of time derivative of the voice signal. This fact is especially obvious in Figure 4, where the time derivatives of the pcm signals are shown. An effective method for locating the starting points of pitch periods is by convolute the time derivative of voice signal with an asymmetric window to obtain an excitation profile function.<sup>11</sup> A well-verified form of the asymmetric window function is Equation (9), shown in Figure 6:

$$w(n) = \sin \frac{\pi n}{N} \left( 1 + \cos \frac{\pi n}{N} \right) \quad -N < n < N. \quad (9)$$

A typical excitation profiles generated by convoluting the asymmetric window Equation (9) with a voice signal is shown in Figure 7. As shown, the results depend on the polarity of the window function. With a correct polarity, the positive peaks of the excitation profile function locate just before the sharp surge of the waveform. If the polarity is opposite, then the positive peaks of the excitation profile function are in the middle of a pitch period. The polarity can be determined by a simple test: If the polarity is correct, the total power of all first halves of the pitch periods is much greater than the total power of all second halves, see Section 5.5.4 of Reference 11. If it is not correct, then the negative peaks are taken as the pitch marks. The size of the window and the threshold of picking up a point can be determined by a brief inspection of the sound signals, see Section 5.4 in Reference 11. This process takes a few milliseconds.

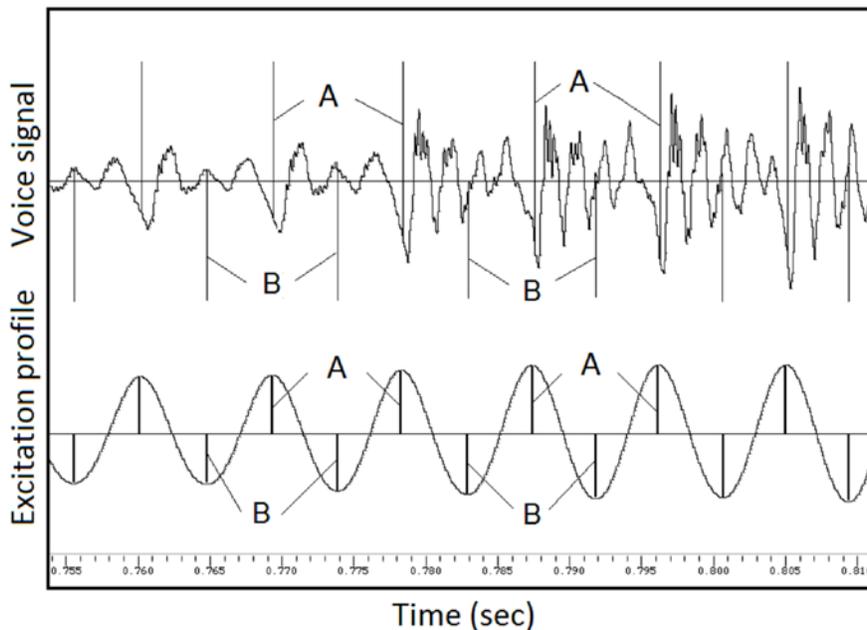


FIGURE 7: Excitation profile function generated by convoluting the voice signal with an asymmetric window function, showing the effect of polarity.

Here we explain the origin of the polarity (positive or negative). Ladefoged's description of the sound of speech as in Equations (6) and (7) is phenomenological. The origin of the sharp pulse that triggers a damped wave is well understood from the correlation of voice signals and the electroglottograph (EGG) signals: it coincides with a glottal closing instant.<sup>10,11</sup> After the glottis is closed, the only available energy source for the speech sound is the kinetic energy of the moving air left in the vocal tract. In the closed phase of the glottal cycle, the speech power decays exponentially due to radiation and friction. An opening of the glottis connects the vocal tract with the porous lungs, and the decay is accelerated. Therefore, the exponential decay waves after Ladefoged is a fairly accurate description of the voice signal. The starting pulse of the speech signal at the glottal closing instant is a negative peak of the perturbation pressure. However, the electrical output from the microphone and amplifier could be either inverted or non-inverted. Therefore, the initial pulse could be either positive or negative. The polarity can be determined by a simple test. Note that the polarity only depends on the microphone and amplifier, independent of the speaker.<sup>11</sup>

With the pitch marks, the sound signals, typically in pcm format, are then segmented into pitch periods. For unvoiced sections, the sound signals are segmented with a constant time interval roughly equal to the average pitch period in the voiced sections.<sup>11</sup>

In general, the pcm values of the two ends of a pitch period do not match. This can be fixed by a minor adjustment of the waveform in each pitch period.<sup>10,11</sup>

## **Fourier Analysis and Display**

The process of Fourier analysis follows the procedure of Hermann, shown by Equations (1) through (4). Initially, the values of the amplitude spectrum are at the integer multiples of the pitch period, Equation (1). Then, the amplitude spectrum is interpolated to a fixed number of points for display, according to the input from the user. The details of the timbre spectrum of each pitch period can be displayed in a separate window. The entire process is now implemented as a software package, the Speech Analyzer. An evaluation version can be accessed through Columbia Technology Ventures, free for academic users.

## **The Speech Analyzer**

To demonstrate the universal usability of the concept and the method, a user-friendly software of pitch-synchronous spectroscopy of speech is developed, which can display the spectrogram and pitch-synchronous spectrum in each pitch period individually. It can handle speech files of any sample rate, mono or stereo, as long as with 16-bit depth in Microsoft .wav format.

The entire software package, including speech samples, is included in a directory, temporarily named *spe*. A user can change the name. The content of the directory is shown in Figure 8(A).

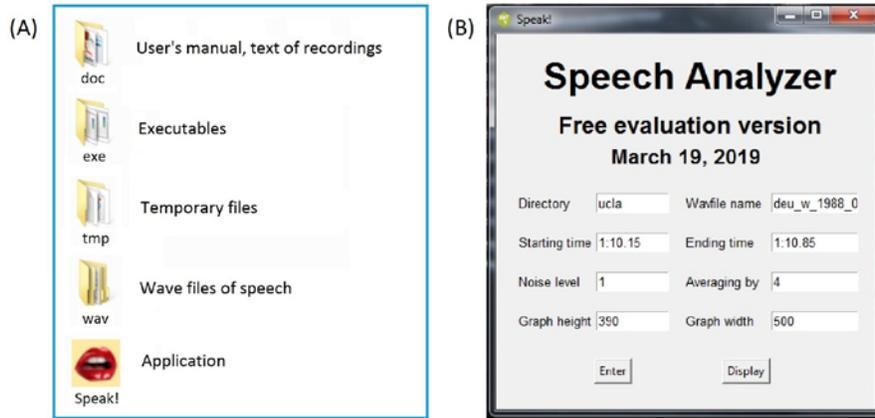


FIGURE 8: Directory structure of the package and the information panel

The subdirectory *doc* contains user's manual and the text of recordings. The subdirectory *exe* contains three executables of the software. The subdirectory *tmp* contains temporary files during the execution. The subdirectory *wav* contains speech signals, in Microsoft wav format. The user can add any wav files into that subdirectory. Currently, it has three subgroups: *ucla* contains recordings of six languages (Arabic, German, French, Dutch, Norwegian, and Portuguese). The file name follows the original file name, except "word-list" is shortened to "w", because the user interface software tcl/tk does not allow hyphen in file names. The subdirectory *eng* contains 10 sentences from the ARCTIC speech-synthesis database from CMU, and the subdirectory *zho* contains 10 sentences from the SpeechOcean Mandarin speech database. By double-clicking the icon of the application *Speak!*, an information window appears, as shown in Figure 8(B).

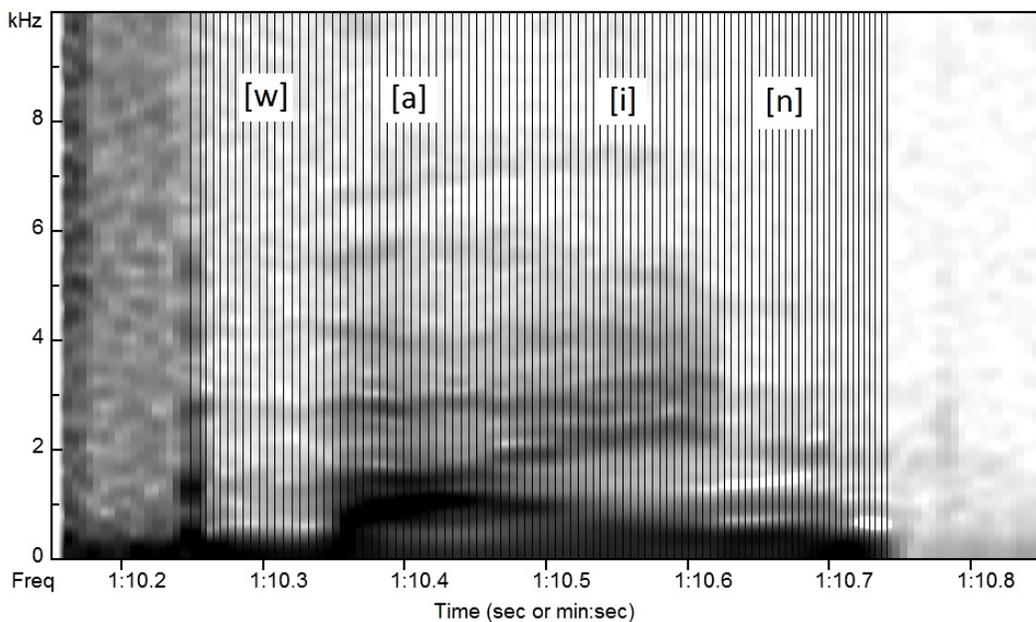


FIGURE 9: Pitch-synchronous spectrogram of the German word Wein.

The fields *Directory* and *Wavfile name* are self-evident. The format of *Starting time* and *Ending time* are *min:sec*; and if the time is less than one minute, it is *sec* only. Because all pitch marks are displayed, and each second has 100-250 pitch periods, to make a good display, the preferred time interval is from 0.25 sec to 1.5 sec. To accommodate noisy recordings, an entry *Noise level* is provided. For clean recordings, the noise level setting is very forgiving. It can be between 1 and 10. For noisy recordings, the accuracy of pitch-period segmentation can be adjusted by setting a number and look at the pitch marks. For example, for some very early recordings in the UCLA USPID collections, a higher number is useful. Because noise exists anyway, and the speaker often have shimmer and jitter, a running average can be applied for the timbre spectrum display. The number of periods to be averaged is set with entry *Averaged by*.

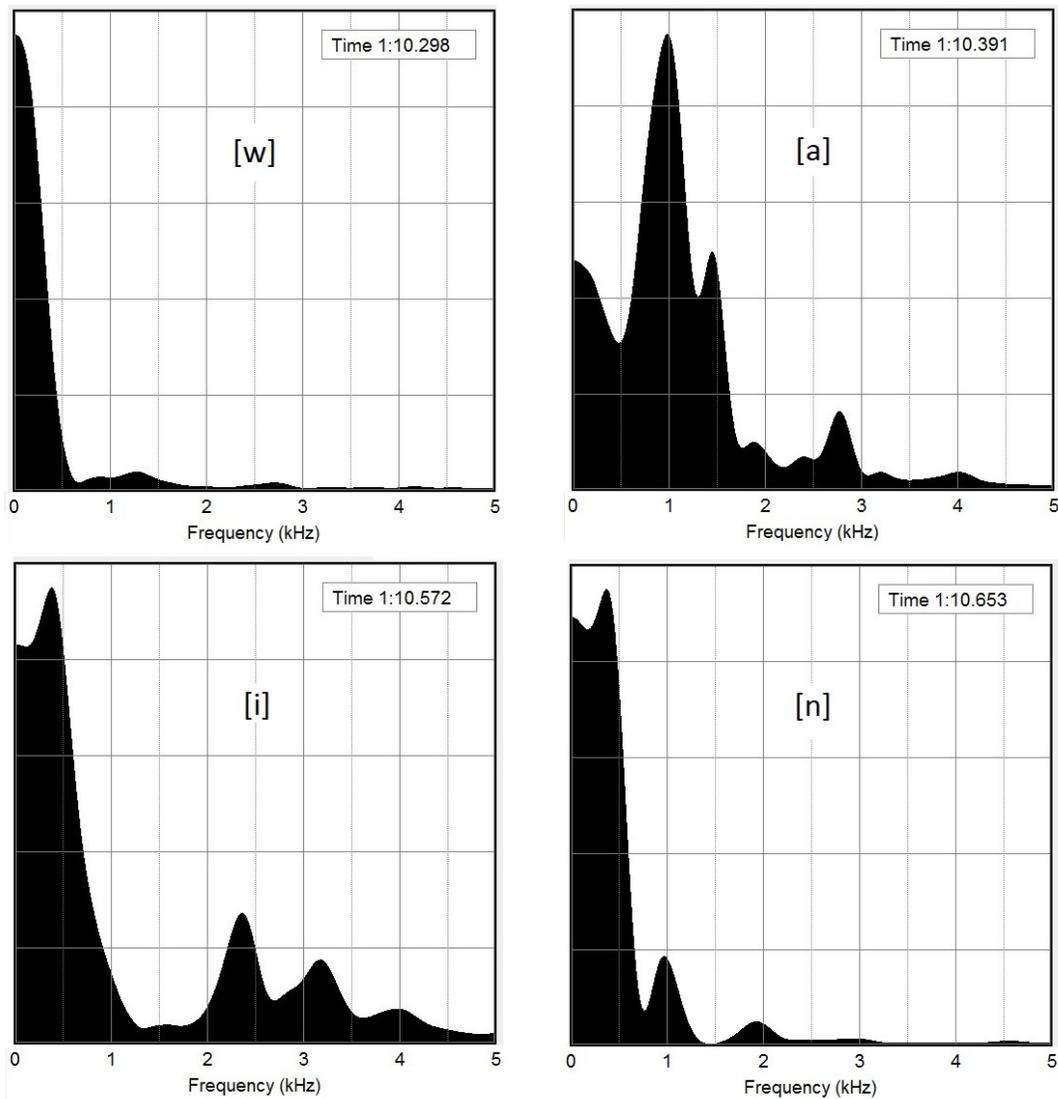


FIGURE 10. Timbre spectra of four sounds in the German word Wein

To choose a file and a time, the speech signal is first displayed using any available software, such as Wavesurfer. You may listen to the sound using Wavesurfer. For example, in the /ucla/deu\_w\_1988\_01.wav, to display the word Wein, the time is set from 1:10.15 to 1:10.85.

By clicking the *Enter* button and then the *Display* button, a pitch-synchronous spectrograph of the speech section appears, as shown in Figure 9.

The height and the width of the graphs can be controlled by *Graph height* and *Graph width* to fit the computer screen. Each vertical line represents a pitch mark, i.e. the starting impulse of a pitch period. The timbre spectrum of each pitch period is displayed as a gray level graph in linear scale to the amplitude, normalized to unity. The dark horizontal bars show the formants. As shown, the pitch information and the timbre information are cleanly separated.

By putting the cursor at any pitch period (time), a mouse click brings a detailed graph of timbre spectrum of that pitch period. A left click brings a narrow-band spectrum, up to 5 kHz. A right click brings a wide-band spectrum, up to 10 kHz. Figure 10 shows four examples. As shown, the depth of information, reproducibility and accuracy are unprecedented.

## Discussions

Pitch-synchronous analysis of speech is not new. After Edison invented the phonograph, using very primitive means and heavily relying on manual work, Ludimar Hermann made Fourier analysis on the waveforms of individual pitch periods of the sound of speech, discovered the basic properties of formants, which has been followed by voice scientists ever since.

Taking advantages of the high-resolution audio acquisition and transmission technology, plus the tremendous computing power unimaginable years ago, following Hermann's basic procedure, much more detailed understanding of the spectrum of speech sound should be gained. Initial experiments on a variety of speech samples showed that the spectroscopic data from the new software module have much better information depth, accuracy and reproducibility.

Although the model of a few formants has done a great deal in speech science, the new generation of spectroscopic data could require a more detailed mathematical representation beyond the formant model. Especially, real-time magnetic resonance imaging of vocal tracts is become a reality. Using numerical integration of the time-dependent wave equation inside the vocal tract, more detailed understanding of voice production mechanism can be achieved. .

## Conclusions

By first segmenting speech signals into natural units -- pitch periods then making Fourier analysis, the richness and accuracy of spectroscopic information is substantially better than the

traditional pitch-asynchronous formant tracking using linear predictive coding method. The richness of the pitch-synchronous spectroscopy of speech suggests a more detailed study of the voice production mechanism beyond a small number of formants, by comparing with the images of vocal tracts from real-time magnetic resonance imaging using numerical integration of the time-dependent wave equation inside the vocal tract.

## Appendix I: Formants According to Ladefoged

According to Ladefoged,<sup>9</sup> the voice of speech consists of a number of decaying waves with a central frequency  $F$  and decay constant  $\beta$ , starting at the instant of a pitch pulse. Mathematically each decaying wave can be represented by Equation (6),

$$s(t) = Ce^{-\beta t} \sin 2\pi Ft \quad 0 < t < \infty. \quad (\text{A1})$$

Here  $t=0$  is the instant of the pitch pulse. The Fourier transform of Equation (A1) is

$$\begin{aligned} S(f) &= C \int_0^{\infty} e^{-2\pi i f t} e^{-\beta t} \sin 2\pi F t dt \\ &= \frac{C}{2} \int_0^{\infty} \left[ e^{-(\beta - 2\pi i(f-F))t} - e^{-(\beta - 2\pi i(f+F))t} \right] dt \\ &= \frac{C}{2} \left[ \frac{1}{\beta - 2\pi i(f-F)} - \frac{1}{\beta - 2\pi i(f+F)} \right]. \end{aligned} \quad (\text{A2})$$

The second term in the last line of Equation (A2) is much smaller than the first term. The amplitude of the Fourier transform is approximately

$$A(f) \propto |S(f)| = \frac{C}{\sqrt{\beta^2 + 4\pi^2(f-F)^2}}. \quad (\text{A3})$$

It is a spectral curve peaks at the formant frequency  $F$ . For simplicity, in Equation (A3), the factor 1/2 is omitted. For a number of decaying waves, the spectrum is approximately

$$A(f) = \sum_{n=1}^N \frac{C_n}{\sqrt{\beta_n^2 + 4\pi^2(f-F_n)^2}}. \quad (\text{A4})$$

In Equation (A4),  $C_n$ ,  $F_n$ , and  $\beta_n$  are the amplitude, frequency, and decay constant of the  $n$ -th formant. Obviously, the decay constant determines the bandwidth of the formant.

## Appendix II: Formant Tracking Based on LPC

Linear prediction theory is an essential component of the theories of signal processing, control, and communication.<sup>12</sup> It was first developed by Wiener during WWII to predict the coordinates of moving targets, and achieved great success.<sup>13</sup> Since 1960s, it was applied to speech signal processing, often called linear predictive coding (LPC).<sup>14,15,3,4</sup> Due to its very wide applications, the mathematics of LPC is well established. In the early years of computers, speed was slow and memory was expensive. The LPC method requires very little computing power (by using e.g. Durbin's recursive algorithm) and the parametric representation is parsimonious.<sup>3</sup> Therefore, it was widely used in speech technology and implemented in speech analysis software packages. However, the information depth, reproducibility, reliability, and accuracy are not good. Here is a brief summary following the notations of Rabiner et al.<sup>3,4</sup>

By representing the digitized speech signal as a time series  $s(n)$ , the LPC model assumes that a signal can be approximately represented by a linear combination of previous signals,

$$s(n) \cong \sum_{k=1}^p \alpha_k s(n-k). \quad (\text{A5})$$

In Equation (A5),  $p$  is the *order* of linear prediction, and the coefficients  $\alpha_k$  are determined by minimizing the square of error  $e(n)$  over a segment of time, defined as

$$e(n) = s(n) - \sum_{k=1}^p \alpha_k s(n-k). \quad (\text{A6})$$

In the z-transform format,<sup>3,4</sup> Equation (A6) has the form

$$E(z) = \left[ 1 - \sum_{k=1}^p \alpha_k z^{-k} \right] S(z), \quad (\text{A7})$$

where  $S(z)$  is the z-transform of the signals,  $s(n)$ , see Equation (A8),

$$S(z) = \sum_{n=-\infty}^{\infty} s(n) z^{-n}, \quad (\text{A8})$$

and  $E(z)$  is the z-transform of the error,  $e(n)$ . Equation (A7) can be reinterpreted that the speech signal is generated by inverse filtering from a source, shown in Equation (A9),

$$S(z) = \frac{U(z)}{1 - \sum_{k=1}^p \alpha_k z^{-k}}, \quad (\text{A9})$$

where  $U(z)$  is the z-transform of the source signals, and  $H(z)$  is the transfer function. The denominator of Equation (A9) is a polynomial, which can be factorized into

$$1 - \sum_{k=1}^p \alpha_k z^{-k} = \prod_{k=1}^p (1 - z_k z^{-1}), \quad (\text{A10})$$

where  $z_k$  are the roots of the polynomial. Because all coefficients  $\alpha_k$  are real, the roots  $z_k$  must be in complex conjugate pairs. For an LPC system of order  $p$ , there should be  $p/2$  pairs of complex roots. In polar coordinates, those roots are

$$z_k = r_k e^{\pm i\theta_k}, \quad (\text{A11})$$

here  $k=1, 2, \dots, p/2$ . The roots have been correlated to the formants,<sup>15</sup> each with frequency

$$F_k = \frac{f_s}{2\pi} \theta_k, \quad (\text{A12})$$

where  $f_s$  is the sample rate. The bandwidth is

$$B_{Fk} = -\frac{f_s}{2\pi} \ln r_k. \quad (\text{A13})$$

Because the radii  $r_k$  are extremely sensitive to minor errors in the coefficients  $\alpha_k$ , usually only the frequencies in Equation (A12) are taken into account. Approximate values of the angles  $\theta_k$  can be obtained from the amplitude of Equation (A10), usually expressed in dB scale,

$$f(\theta) = -20 \log_{10} \left| 1 - \sum_{k=1}^p \alpha_k e^{-ik\theta} \right|. \quad (\text{A14})$$

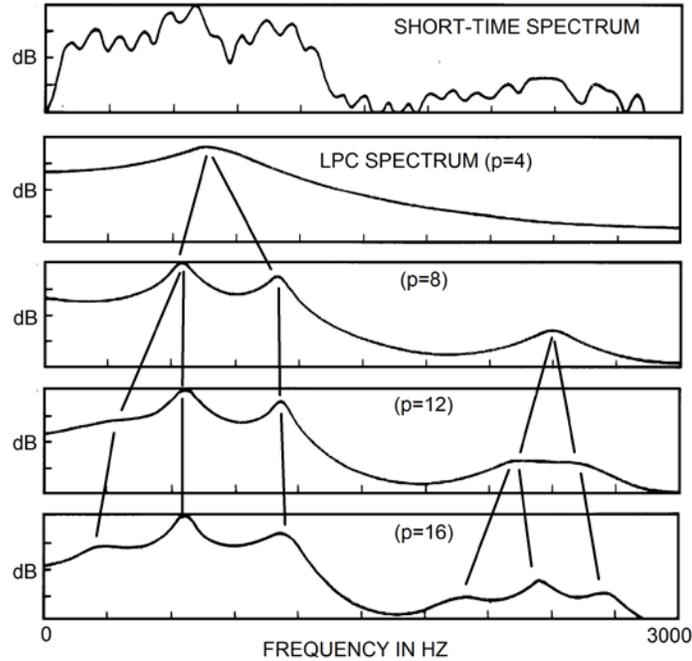


FIGURE A1. Formant frequencies determined by LPC spectrum.

Adapted from page 435 of Ref. 3 and page 111 of Ref. 4.

Figure A1 shows an example of Equation (A14) (see page 435 of Ref. 3 and page 111 of Ref. 4). The peaks in the function Equation (A14) are assumed to be the estimates of formant frequencies through Equation (A12), thus the number of formants equals one half of  $p$ . As shown, when order  $p$  changes, the frequencies of the estimated formants also changes. Often, one estimated formant at a lower order splits into two or three estimated formants at higher orders.

The process and properties of LPC are discussed in detail on pages 112-115 of Reference 4. Many of the additional peaks in the LPC spectrum function for larger  $p$  in Equation (A14) are generally irrelevant ones. They do not reflect the genuine formants of the voice signal. Rather, those peak frequencies are individual harmonics of the pitch frequency  $f_0$ . It is very difficult to distinguish true formant frequencies from irrelevant peak frequencies in the LPC spectrum.

There are more arbitrariness in the process of extracting the LPC coefficients  $a_k$  in Equation (A5). First, the speech signal is segmented into overlapping frames, typically of a width of 25 msec and a shift of 10 msec, and multiplied by a window function, typically a Hamming window.<sup>3,4</sup> Then the signal goes through a *pre-emphasis* step.<sup>3,4,14</sup> The higher-frequency components of the signal are pre-emphasized preferentially. For example, the signal component at 3 kHz can be pre-emphasized by 10 folds over the signal component at 300 Hz. Otherwise the formants of higher frequencies would not show up. However, by pre-emphasizing the high-frequency components, for speech signals factually only have low-frequency formants, such as [u], [m], [n] etc. spurious formants in the high-frequency range would appear. The size and shift of the windows and the multiplicative window functions also affects the measurement results.

Furthermore, as shown in Equation (A4) in a previous Section, each formant should have three parameters: central frequency, intensity, and bandwidth. Intensity is very important. For example, the different of [i] and [u] depends on the relative intensity of the low-frequency formants and the high-frequency formants. By gradually decrease the intensity of the high-frequency formants, the vowel [i] gradually becomes vowel [u]. As apparent from Equations (A12) and (A13), the LPC method only provides two parameters from each formant, the central frequency and the bandwidth. No intensity can be predicted by the LPC method. Therefore, the LPC parameters would not able to distinguish [i] and [u] in certain cases.

In speaker identification and verification, especially in forensic phonetics, the accuracy and objectiveness of measurement methods are of paramount importance. An extensive study of the accuracy of formant determination regarding to forensic research was published as a PhD thesis by Harrison in 2013.<sup>6</sup> As expected, a severe inadequacy of the LPC method is the artifact of the order  $p$  on the estimated frequencies of formants.<sup>6</sup>

## REFERENCES

1. Ladefoged P and Maddieson I, *The Sounds of the World's Languages*, Blackwell Publications, Oxford, UK, 1996.
2. Ladefoged P, *Course in Phonetics*, Harcourt Brace Jovanovich, Inc., New York, 1982.
3. Rabiner LR and Schafer RW, *Digital Processing of Speech Signals*, Prentice-Hall, New Jersey, 1978.
4. Rabinar L and Juang BH, *Fundamentals of Speech Recognition*, Prentice-Hall, New Jersey, 1993.
5. Beigi H, *Fundamentals of Speaker Recognition*, Springer, Berlin, 2011.
6. Harrison PT, *Making Accurate Formant Measurements*, PhD Thesis, University of York, Language and Linguistics Science, 2013.
7. Hermann L, Phonophotographische Untersuchungen VI. Nachrag zur Untersuchung der Vocalcurven Pflüger Archiv, now European Journal of Physiology. Vol 43, page 264, 1894.
8. Scripture EW, *The Elements of Experimental Phonetics*, Charles Scribner's Sons, London, 1902.
9. Ladefoged P, *Elements of Acoustic Phonetics*, Second Edition, University of Chicago Press, Chicago, 1996.
10. Chen CJ and Miller DG, *Pitch-Synchronous Analysis of Human Voice*, Journal of Voice, 2019.
11. Chen CJ, *Elements of Human Voice*, World Scientific Publishing 2016.
12. Strobach P, *Linear Prediction Theory*, Springer Verlag, Berlin, 1990.
13. Wiener N, *Extrapolation, Interpolation, and Smoothing of Stationary Series*, MIT Press 1949.
14. Makhoul J, *Linear Prediction, a Tutorial Review*, Proc. IEEE, 63, 561-580 (1975).
15. Markel JD and Gray AH Jr, *Linear Prediction of Speech*, Springer, Berlin, 1976.
16. Open webpage [innovation.columbia.edu/technologies/CUI8357](http://innovation.columbia.edu/technologies/CUI8357), then follow the instructions.