Pitch-Synchronous Biomarkers for Voice-Based Diagnostics

C. Julian Chen *

Posted Date: 29 August 2025

doi: 10.20944/preprints202508.2098.v1

Keywords: voice-based diagnostics; jitter; shimmer; preventive madicine; timbre vector

 doi:10.20944/preprints202508.2098.v1

Article

Pitch-Synchronous Biomarkers for Voice-Based Diagnostics

C. Julian Chen

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York 10027; jcc2161@columbia.edu

Abstract

Background: Voice analysis combined with artificial intelligence (AI) is rapidly becoming a vital tool for disease diagnosis and monitoring. A key issue is the identification of vocal biomarkers, that are quantifiable features extracted from voice to assess a person's health status or predict the likelihood of certain diseases. Currently, the biomarkers are extracted using pitch-asynchronized methods that need improvement. **Methods:** Based on the timbron theory of voice production, a pitch-synchronous method of vocal biomarker extraction is proposed and tested. **Results:** As examples, the methods are applied on the ARCTIC speech databases published by Carnegie Melon University, and the Saarbrucken voice database for voice diagnostics. From the ARCTIC database, a complete set of formant parameters and timbre vectors for all US English monophthong vowels are presented. The timbre distances among all US English monophthong vowels are presented, showing the richness and accuracy of information contained in those biomarkers. By applying pitch-synchronous methods on the voice recordings in the Saarbrucken voice database, accurate and reproducible measurements of timbre vectors, jitter, shimmer, and spectral irregularity are shown. Furthermore, methods of detecting glottal closing instants from voice signals are discussed and demonstrated. **Conclusions:** The biomarkers extracted using pitch-synchronous analysis contain abundant, accurate, objective, and reproducible information from the voice signals that could improve the usability and reliability of voice-based diagnostics.

Keywords: voice-based diagnostics; jitter; shimmer; preventive madicine; timbre vector

1. Introduction

Voice analysis combined with artificial intelligence (AI) is rapidly becoming a vital tool for disease diagnosis and monitoring.[1–7] For many decades, voice-based diagnostics has been applied in otorhinolaryngology to study voice disorders.[1] Because human voice is produced by a combination of several organs, voice-based diagnostics is shown to be applicable also to other medical disciplines.[2] For example, in neurology to diagnose and monitor dementia and Parkinson's disease,[3] in pulmonology for respiratory diseases,[2,4] in oncology for early detection of many types of cancer including laryngeal cancer, throat cancer, oral cancer, and lung cancer.[5] Recently, it was also found that voice-based diagnostics can be effectively applied to detect and monitor heart failures.[6]

Voice-based diagnostics have many benefits. It is non-invasive and can be more comfortable than traditional diagnostic procedures for patients. It enables early detection. It is easily accessible and with reduced costs. Especially, patients could monitor their health by submitting voice samples remotely to a healthcare provider, enabling ongoing evaluations and personalized care.

A key issue currently under intensive study is the identification of vocal biomarkers, that are quantifiable features extracted from voice signals to assess a person's health status or predict the likelihood of certain diseases.[2,8,9] Currently, the traditional methods for feature extraction in speech technology using pitch-asynchronous analysis methods are used.[10,11] The following biomarkers are often utilized.[6,13]

 doi:10.20944/preprints202508.2098.v1

Mel-frequency cepstral coefficients (MFCCs), the coefficients derived by computing the cepstrum of a log-magnitude spectrum based on the Mel scale of voice perception.[14,15] It is a widely used biomarker.

Linear frequency cepstral coefficient (LFCCs), the coefficients derived by using a linear scale of frequency.

The first two formants F1 and F2, which are the two lowest resonance frequencies of the vocal tract.

Linear prediction coefficients (LPC).[30,31] It is widely used in speech technology and voice analysis but not frequently used in voice-based diagnostics.[1]

All the above biomarkers are extracted using a pitch-asynchronous analysis method. During the extraction process, a lot of vital information is lost. As we will show in this paper, based on a better understanding of voice production and a pitch synchronous parameterization method, better biomarkers can be extracted.

According to the physiology of human voice production, as presented in Section III,[18–21] human voice is generated a pitch period at a time, staring at each glottal closing instant (GCI). The elementary sound wave triggered by each GCI, called a timbron, contains full information on the timbre. Continuous voice is generated by a superposition of a sequence of timbrons triggered by a series of glottal closing events. According to the timbron theory of voice production, the time difference between two adjacent GCIs defines the pitch period, and the sound waveform in each pitch period contains full information on the timbre.

To extract more information from voice signals, a pitch-synchronous analysis method is developed.[18–21] By using that method, pitch information and timbre information are cleanly separated. On average, the pitch period is 8 msec for men, and 4 msec for women. Therefore, for every 4 to 8 msec, a complete and accurate set of information on the timbre and pitch of human voice can be obtained. The biomarkers extracted using a pitch-synchronous analysis method contain abundant, accurate, objective, and reproducible information from the voice signals that could improve the usability and reliability of voice-based diagnostics.

The organization of the article is as follows.

In Section II, the deficiencies of the traditional pitch-asynchronous analysis methods are analyzed. It was developed in the middle of the 20th century, when the computing power was low and computing languages were underdeveloped. A lot of vital information is lost during the extraction process.

In Section III, the timbron theory of voice production, a modern version of the transient theory, is presented as a logical consequence of the temporal correlation of the voice signals and the simultaneously acquired electroglottograph (EGG) signals. According to the timbron theory, the time difference between two adjacent glottal closing instants (GCIs) defines the pitch period (the inverse of which is the pitch frequency), and the sound waveform in each individual pitch period contains full information on the timbre of the voice.

In Section IV, a pitch-synchronous method of voice analysis is presented. By applying the pitch-synchronous analysis method to a standard US English speech corpus, the formant parameters of all US English monophthong vowels are measured, see Section V. The correctness of the formant parameters is tested by voice synthesis, using a program appended to the article.

The format parameters are not the best biomarkers for voice-based diagnostics. In Section VI, the definition of timbre vectors together with the method of extraction, is presented. From a standard US English speech corpus, the timbre vectors for all monophthong vowels are presented. Especially, the timbre distances among all US English monophthongs are presented, showing the reliability and accuracy of the method.

In Section VII, the methods of finding jitter, shimmer, and spectral irregularities are presented. It is based on the recordings in the Saarbrucken voice database, showing the effectiveness of the pitch-synchronous analysis method and the usefulness of timbre vectors.

 doi:10.20944/preprints202508.2098.v1

In Section VIII, the methods for detecting GCIs from voice signals are presented. Based on the reference GCIs from the EGG, the accuracy and usefulness of the methods of detecting GCIs from voice signals is discussed. The values of simultaneously acquired EGG signals are also discussed. Section IX presents results and discussions.

2. Pitch-Asynchronous Analysis Methods

The evolution of voice analysis methods was closely related to the evolution of computing power. The methods of voice analysis were advanced and also limited by the available computing power at various stages of time. In the 1960s and 1970s, voice processing was implemented on mainframe computers with very low processing power. Accordingly, LPC, that requires low computing power, emerged at that time.[30,31] In the 1980s, PCs were born and evolved quickly. Fast Fourier transform (FFT) was applied to voice processing. MFCC was invented in 1980,[14] and then replaced LPC as the leading parametrization method for speech recognition.

The advantages of pitch-synchronous analysis for speech were recognized and studied in the 1970s and 1980s.[16,17] Nevertheless, it requires much higher processing power and did not become a mainstream.

The general procedure of pitch-asynchronous analysis is shown in Figure 1. Voice signal is typically blocked into frames with a duration of 25 msec and a shift of 10 msec, then multiplied with a window function, typically a Hamming window.[10,11] Because the windows are not aligned with pitch periods, pitch and timbre information are mixed. Even the voiced and unvoiced sections can be mixed, see Figure 1(A) and (B).

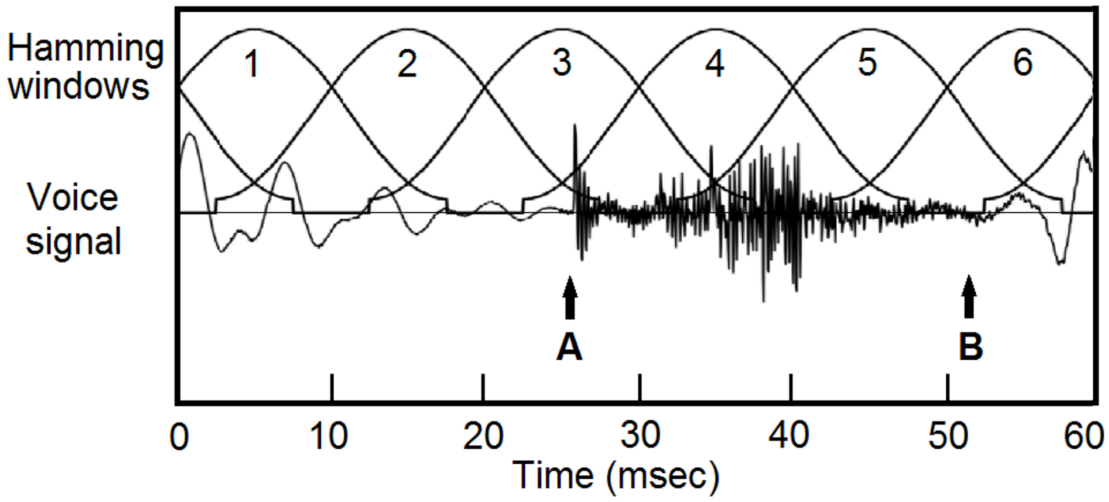

Figure 1. Pitch-asynchronous processing of voice signals. Speech signal is typically blocked into 25 msec frames with 10 msec shift, then multiplied with a window function, typically a Hamming window. Pitch boundaries are missed, and even voiced and unvoiced sections are mixed, see A and B.

The use of frames of a fixed size is also mandated by the early fast Fourier transform (FFT). Traditionally, the size of source data for FFT must be a power of 2, typically 256 or 512. Because the windows are not aligned with the pitch boundaries, the output of FFT, the amplitude spectrum, is a mix of pitch periods and timbre properties, see Figure 2(A). The features are dominated by the overtones of the fundamental frequency, with an envelop associated with the formants.

 doi:10.20944/preprints202508.2098.v1

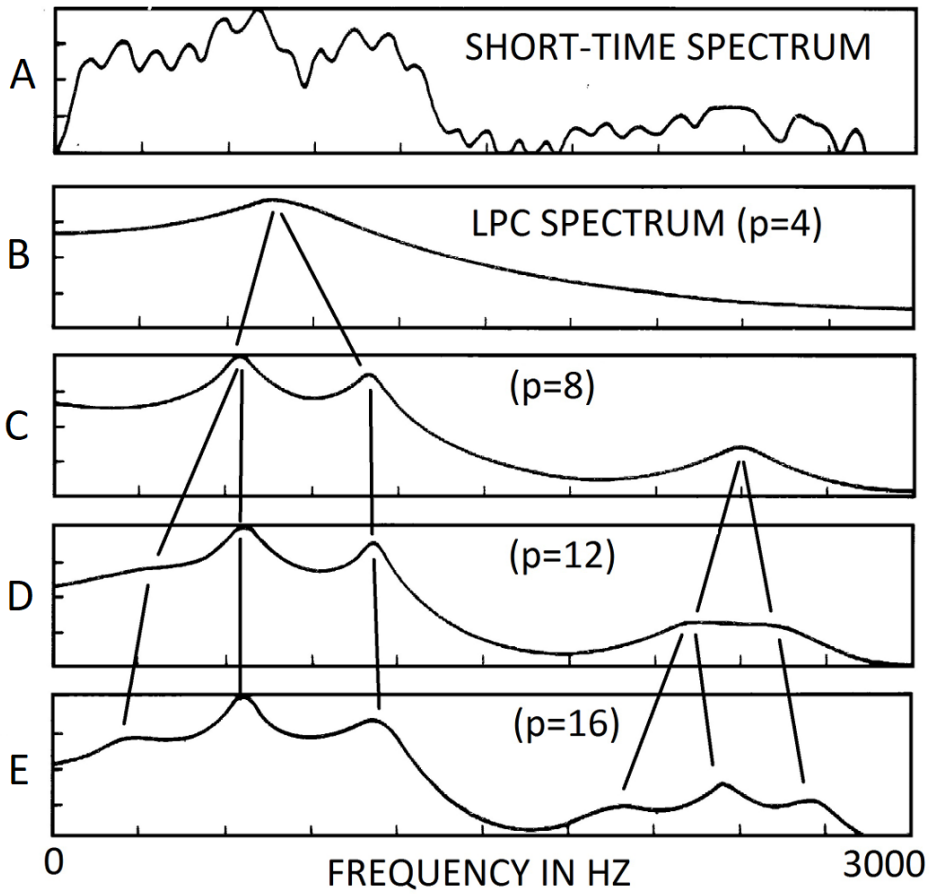

Figure 2. Dependence of formant frequencies on order p of LCP. To measure formants using LPC, the order of transfer function p must be entered first. The number and frequencies of formants depend on p. After Fig 8.18 on page 438 of Rabiner and Schafer [10].

The raw amplitude spectrum cannot be used as a biomarker. In the 1960s and 1970s, to extract usable features from the spectrum, the method of linear prediction coefficients (LPC) was introduced.[30,31] The short-time spectrum, showing in Figure 2(A), is approximated by an all-pole transfer function of order p. LPC is often applied to extract formants. Nevertheless, for each formant, the LPC method can only provide the frequency and the bandwidth. Level, an important parameter, is missing. Furthermore, the results of the frequencies using LPC method depend on an input parameter, the order p of the transfer function, see Figure 2. For example, with $p = 4$, the LPC analysis generates only one formant. With $p = 8$, three formants are shown. The one formant resulting from $p = 4$ is not present. Therefore, that method is neither reproducible nor objective [10,31].

To extract timbre information from the short-time spectrum, in 1980, the method of mel-frequency cepstral coefficients (MFCC) was invented [14] and gradually replaced LPC to become the most used parameters for speech recognition.[11] To separate pitch information and timbre information, it bins the amplitude spectrum using Mel-scaled triangular windows, to accommodate non-linear human perception of frequencies, as shown in Figure 3. Because the number of bins is limited, typically between 10 and 15, the frequency resolution is low.

By comparing with the rich information contained in voice signals available by using a pitch-synchronous analysis method, the MFCC has the following deficiencies, especially regarding voice-based diagnostics:

First, the quantity of feature vectors. Using the pitch-synchronous analysis method, a complete set of information can be obtained from each pitch period, which is roughly 8 msec for men and 4 msec for women. By using a fixed window size, typically 25 msec, the number of feature vectors is dramatically reduced.

 doi:10.20944/preprints202508.2098.v1

Second, the quality of the feature vectors. To extract timbre information from a mix of pitch periods and formants, a binning procedure is taken, see Figure 3. The number of bins is typically 10 to 15. The frequency resolution is seriously limited. In contrast, by using a pitch-synchronous analysis method, the frequency resolution is virtually unlimited. It is practically determined by the quality of the original voice signals.

Third, objectiveness of the results. The choice of window size and the process window function has a dramatic effect on the final results. Therefore, the MFCC coefficients are not objective.

Finally, reproducibility of the results. For example, a reduction or addition of silence changes the positions of the processing windows, and the MFCCs become different, which could affect the results of diagnostics.

By using a pitch-synchronous analysis method as shown in this paper, all those problems are resolved. Not only the quantity and quality of results are higher, but also are objective and reproducible. The results, for example the timbre vector, are extracted from each pitch period. The processing frame boundaries are determined by the voice itself, not imposed manually.

The recent progress of computing hardware and programming languages enables the implementation of pitch-synchronous analysis methods. Coming to the 21st century, the evolution of computer power has been accelerated to be doubled in a few months. Owing to the growing computing power, Python programming language, although executes much slower than C++, became the most widely used programming language because of its readability and maintainability. Using the Numpy module, FFT can be applied to an array of any size, not limited to power of 2. It greatly streamlines the implementation of pitch synchronous analysis method.

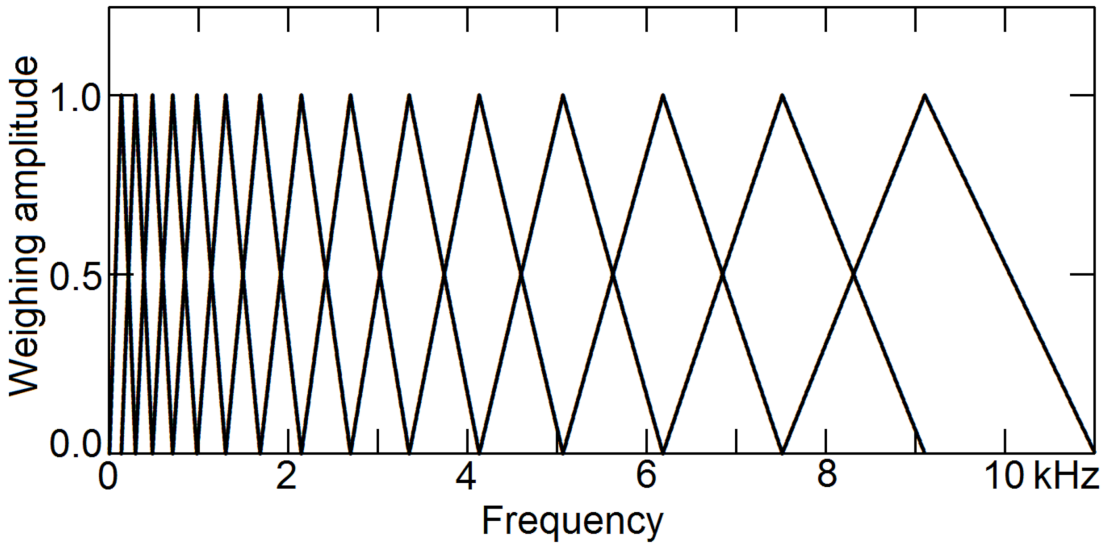

Figure 3. Mel-scaled triangular windows. The human perception of frequency scale is non-linear. To build MFCC, a set of rectangular windows of non-linear scale is multiplied to the spectrum, to generate a set of coefficients.

3. Physiology of Voice Production

In order to define a better set of biomarkers and correlate the test results to a person's health status and pathological conditions, a correct understanding of the physiology of voice production is necessary. Here we briefly outline a modern understanding of the physiology of human voice production, timbron theory.[18–21]

The timbron theory of human voice production is a logical consequence of the observed temporal correlation of the voice signals and the simultaneously acquired electroglottograph (EGG) signals.[18–20] EGG was invented by French physiologist Philip Fabre in 1957.[25,26] Since EGG is non-invasive and contains valuable information on the physiology of voice, it has become widely used.[22–24] A huge volume of speech recordings with simultaneously acquired EGG signals have been collected and

 doi:10.20944/preprints202508.2098.v1

published. The ARCTIC database, published by Carnegie Melon University, is a good example.[27] For the study of voice-based diagnostics, the Saarbrucken voice database, with 2250 voice recordings, available for free, was also collected with simultaneously acquired EGG signals.[1,28]

The working principle of EGG is shown in Figure 4. Two electrodes are pressed against the right side and the left side of the neck, near the vocal folds. A high-frequency electrical signal, typically 200 kHz, is applied to one electrode, and an AC microammeter is placed on the other electrode, to measure the electrical conductance between the two electrodes. (a), while the glottis is open, the conductance is lower, and the ac current is smaller. (b), while the glottis is closed, the conductance and the ac current is higher. The glottal closing instants (GCIs) can be determined accurately.[22–24]

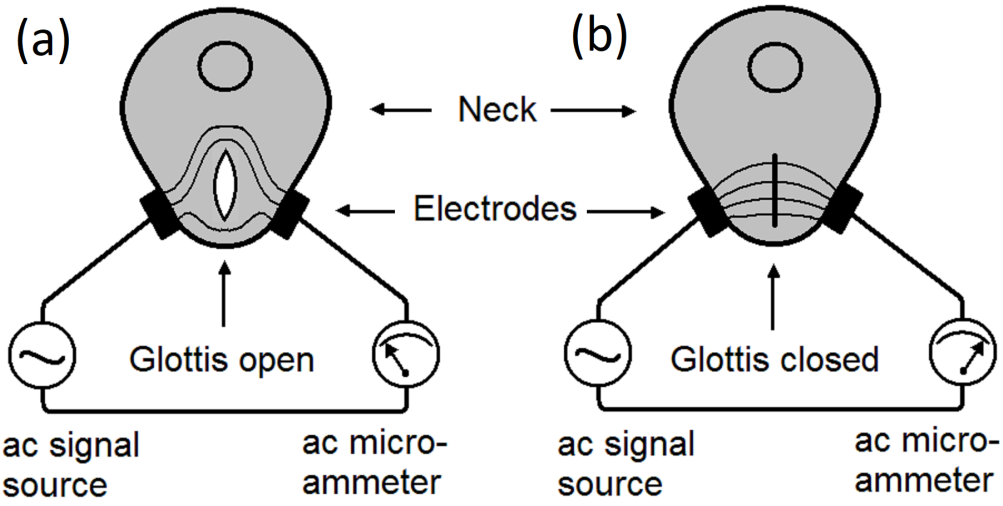

Figure 4. A schematic of the electroglottograph (EGG). A high-frequency ac signal is applied on one side of the neck. An ac microammeter is connected to an electrode placed on the other side of the neck. (a), an open glottis reduces ac current. (b), a closed glottis increases the ac current. Glottal closing instants can be accurately detected.

A ubiquitous experimental fact was observed in the temporal correlation between the voice signal and EGG signal,[18–20] see Figure 5. Right after a GCI, (a), a strong impulse of negative perturbation pressure emerges. During the closed phase of glottis, (b), the voice signal is strong and decaying. In the glottal open phase, (c), the voice signal further decays and becomes much weaker. Immediately before a GCI, (d), there is a peak of positive perturbation pressure. The next GCI starts a new decaying acoustic wave, superposing on the tails of the previous decaying acoustic waves. The elementary decaying wave started at a GCI is determined by the geometry of the vocal tract, containing full information on the timber, thus called a timbron. The pitch period is defined as the time interval of two adjacent GCIs. The observed waveform in each pitch period contains the starting portion of the timbron in the current pith period and the tails of the timbrons of previous pitch periods, therefore contains full information on the timber of the voice.

 doi:10.20944/preprints202508.2098.v1

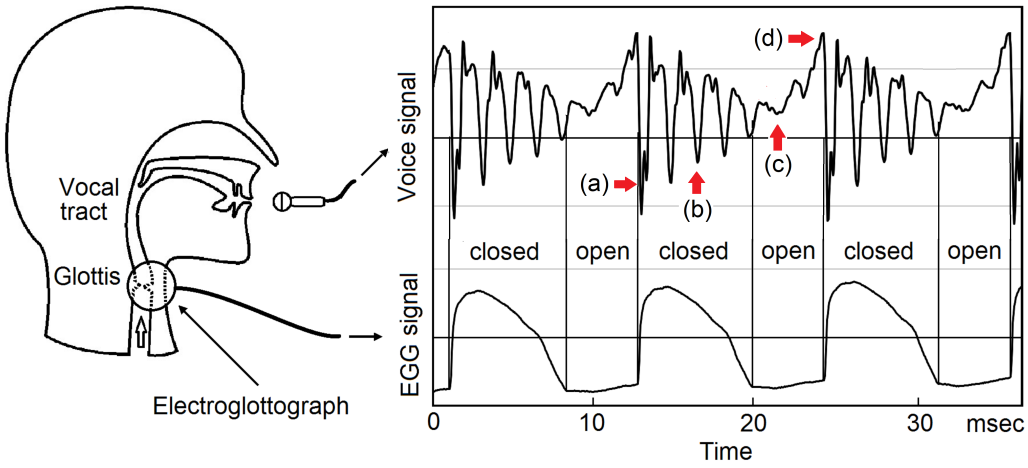

Figure 5. Temporal correlation between the voice signal and the EGG signal. (a). Immediately after a glottis closing, there is a strong impulse of negative perturbation pressure. (b). In the closed phase of glottis, the voice signal is strong and decaying. (c). The voice signal in the open phase of the glottis is much weaker. (d). Immediately before a glottal closing, there is a peak of positive perturbation pressure, representing the glottal flow.

The acoustic process of generating an elementary sound wave characteristic to the vocal tract geometry is shown is shown in Figure 6. During the glottal open phase, step (0), air flows continuously from the trachea to the vocal tract. No sound is produced. At an abrupt glottal closing instant, step (1), the airflow is suddenly blocked. However, the airflow in the vocal tract keeps moving due to inertia. A negative-pressure wavefront is generated, which propagates with the speed of sound towards the lips, see (2). Then, the wavefront reflects from the lips and propagates towards the glottis, see (3) and (4). Because the glottis is closed, the wavefront is again reflected and propagates toward the lips. A complete formant cycle takes four back-and-forth propagations of the acoustic wave over the length of the vocal tract at the speed of sound. The time of a complete cycle is

$$T = \frac{4L}{c}, \tag{1}$$

here c is the speed of sound, 350 m/sec, and L is the length of the vocal tract. For male speakers, $L \approx 17.5$ cm. One has $T = 2$ msec, or a frequency of $F = 0.5$ kHz. If the vocal tract is a lossless tube of uniform cross section, a *square wave* of 0.5 kHz is produced. The Fourier series of a square waves is

$$x(t) = C\left[\sin 2\pi Ft + \frac{1}{3}\sin 6\pi Ft + \frac{1}{5}\sin 10\pi Ft + ...\right], \tag{2}$$

which can be observed as a series of *formants* centered at frequencies 0.5 kHz, 1.5 kHz, 2.5 kHz, and so on.

 doi:10.20944/preprints202508.2098.v1

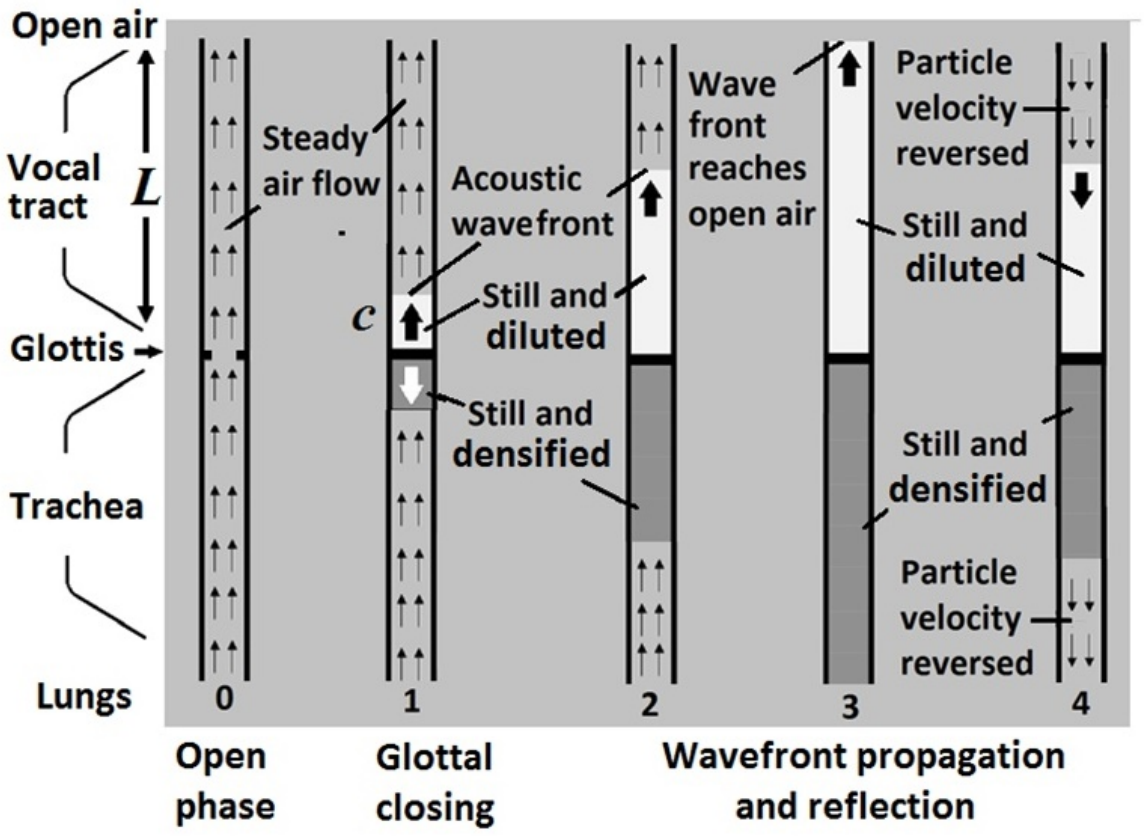

Figure 6. Mechanism of timbron production. (0), while the glottis is open, there is a steady airflow from the lungs to the vocal tract. No sound is produced. (1), a glottal closing blocks the air flow from the trachea to the vocal track. It creates a negative-pressure wavefront moving towards the lips at the speed of sound and resonates in the vocal tract to become a timbron. See references [18–20].

In general, loss of energy is unavoidable, and the shape of the vocal tract is different from a uniform tube. The waveform of an elementary sound wave triggered by a glottal closing with N formants is

$$\begin{aligned} x(t) &= \sum_{n=0}^{n<N} C_n e^{-2\pi B_n t} \sin 2\pi F_n t, \quad t \geq 0; \\ x(t) &= 0, \qquad\qquad\qquad\qquad\quad t < 0. \end{aligned} \tag{3}$$

Here the constants C_n is the amplitude, related to the level of the formant; B_n is the bandwidth width, and F_n is the central frequency of the formant. Note that before a glottal closing, $t = 0$, there is no acoustic signal. The glottis is open during the open phase. Nevertheless, the area of the glottis is much smaller than the mouth and the tissue surface of the vocal tract. The effect to the evolution of the decaying wave is minor. Therefore, the decaying wave continues until it disappears.

If there is a series of glottal closings with a constant time shift, or a constant pitch frequency, the next glottal closure starts a new decaying wave superposed on the tails of the previous decaying waves, see Figure 7. As shown, for different values of time shift, vowels with different pitch frequencies, for example, voice of F_0 = 125 Hz, 100 Hz, and 75 Hz can be produced.

 doi:10.20944/preprints202508.2098.v1

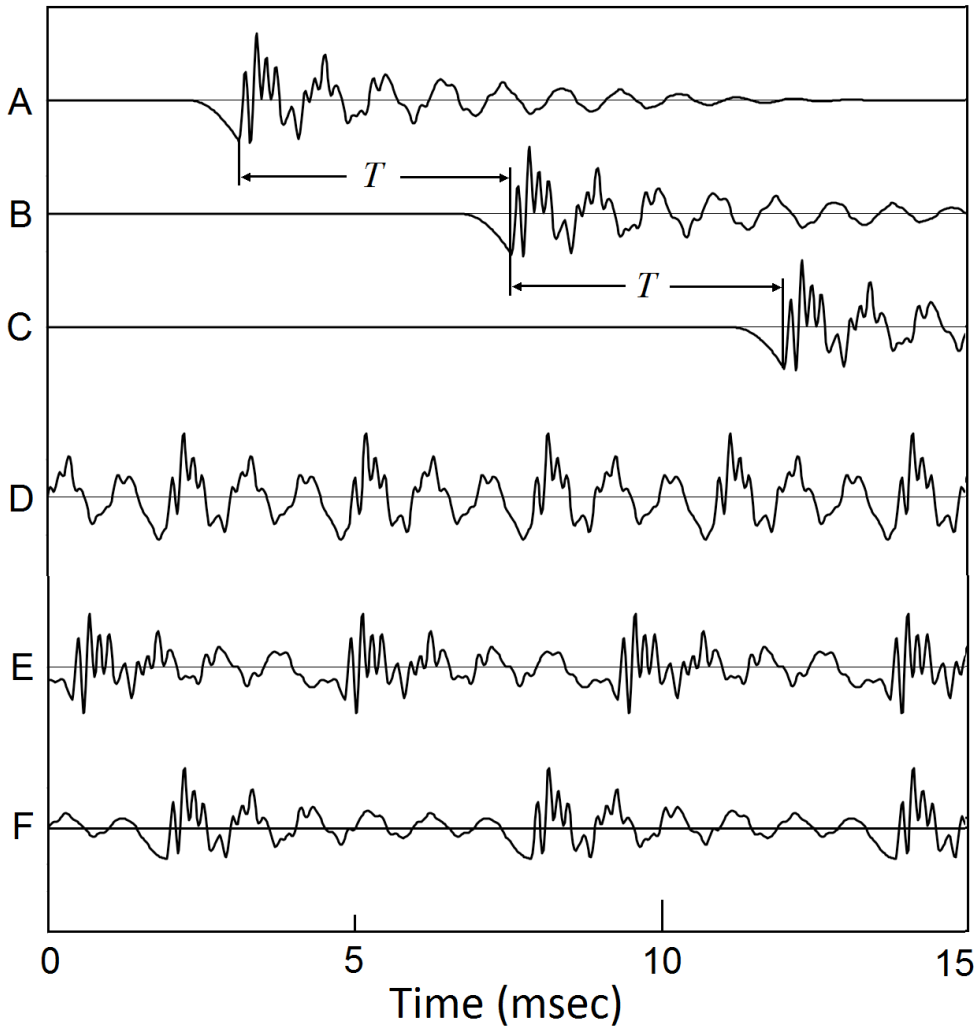

Figure 7. Superposition of timbrons to become a sustaining voice. (a) through (c), with a series of glottal closings, the timbrons started from those GCIs superpose over each other to become sustaining voice. Each new timbron is adding to the tails of the previous timbrons. Different repetition rate of the GCIs make voice of different pitch frequencies. (d), F_0 = 125 Hz. (e), F_0 = 100 Hz. (e), F_0 = 75 Hz.[18–20]

The elementary wave triggered by a glottal closing is formed by the shape of the vocal tract, which contains full information on the instantaneous timbre. It is termed a *timbron*.[18–20] The fact that a timbron is zero before a glottal closing at $t < 0$ has interesting consequences, which can be proved with mathematical rigor.[20]

First, the waveform in each individual pitch period is a superposition of the timbron in the current pitch period and *the tails of the timbrons of the previous pitch periods*. This makes the elementary voice waveform to exhibit a maximum time-inversion asymmetry.

Second, according to a well-known mathematical theorem in the analysis of complex variables, the dispersion relation or the Kramers-Kronig relation, the *phase spectrum is completely determined by the amplitude spectrum*. Therefore, the amplitude spectrum contains full information on the entire waveform.

Third, in computing the Fourier transforms of the waveform in each pitch period, *the starting point of waveform has no effect on the Fourier transform*. It greatly simplifies the computation of the Fourier transform.

 doi:10.20944/preprints202508.2098.v1

4. Pitch-Synchronous Analysis of Voice

As a demonstration of the pitch-synchronous method of extracting biomarkers from the voice signals, we use the ARCTIC database for speech science published by the Language Technologies Institute of Carnegie-Melon University.[27] Released in 2004, it became a standard corpus for speech science and technology. For each speaker, it has 1132 sentences carefully selected based on the requirement of phonetic balance from out-of-copyright texts of Project Gutenberg. The advantages of using this corpus as the first test are: The recordings were carefully collected under well-controlled conditions. It is phonetically labeled according to the ARPABET phonetic symbols.[29] It has simultaneously acquired EGG signals. In this paper, the voice of an US English speaker bdl is used. Note that some of the speech databases for voice-based diagnostics, such as the Saarbrucken voice database, also has simultaneously acquired EGG signals.[1,28] For voice recordings without EGG signals, GCI can be extracted from voice signals. We will discuss this issue in Section VII.

Figure 8 shows a small section of sentence a0329 from speaker bdl including the time derivative of EGG signal. The dEGG/dt curve shows sharp peaks at GCI points. Based on those GCI points, a series of segmentation points are generated. The voice signal is then segmented into pitch periods.

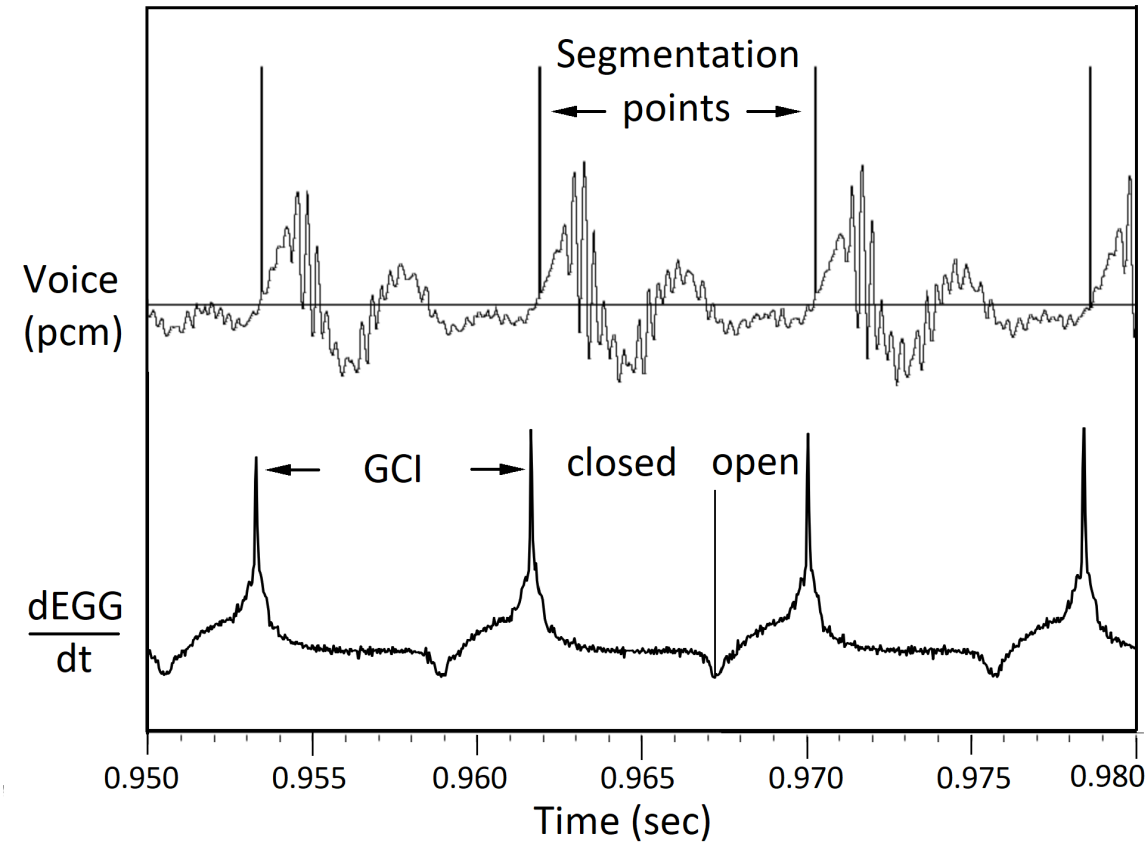

Figure 8. Segmenting voice signals into pitch periods. A section of sentence a0329. At each glottal closing instant (GCI), the dEGG/dt signal shows a sharp peak. Using those GCI points as segmentation points, the voice signal can be segmented into pitch periods.

Because in general, right after segmentation, the value of the starting value A of the voice signal in a pitch period does not equal to the ending value C, an *ends-matching procedure* is executed, see Figure 9. For example, the original waveform in a pitch period is AC, an array **wavin** with 125 floating-point numbers. By taking 10 more points as CD, an array **ext** with 10 floating-point numbers is formed. Using the Python code,

```
for n in range(10):
    x[n] = 0.1*float(n)
```

 doi:10.20944/preprints202508.2098.v1


```
wavin[n] = wavin[n]*x[n]+ext[n]*(1-x[n])
```

the beginning section of the pitch period AB is replaced by the new section A'B. The ends are matched, and the curve is completely continuous, see Figure 9(b).

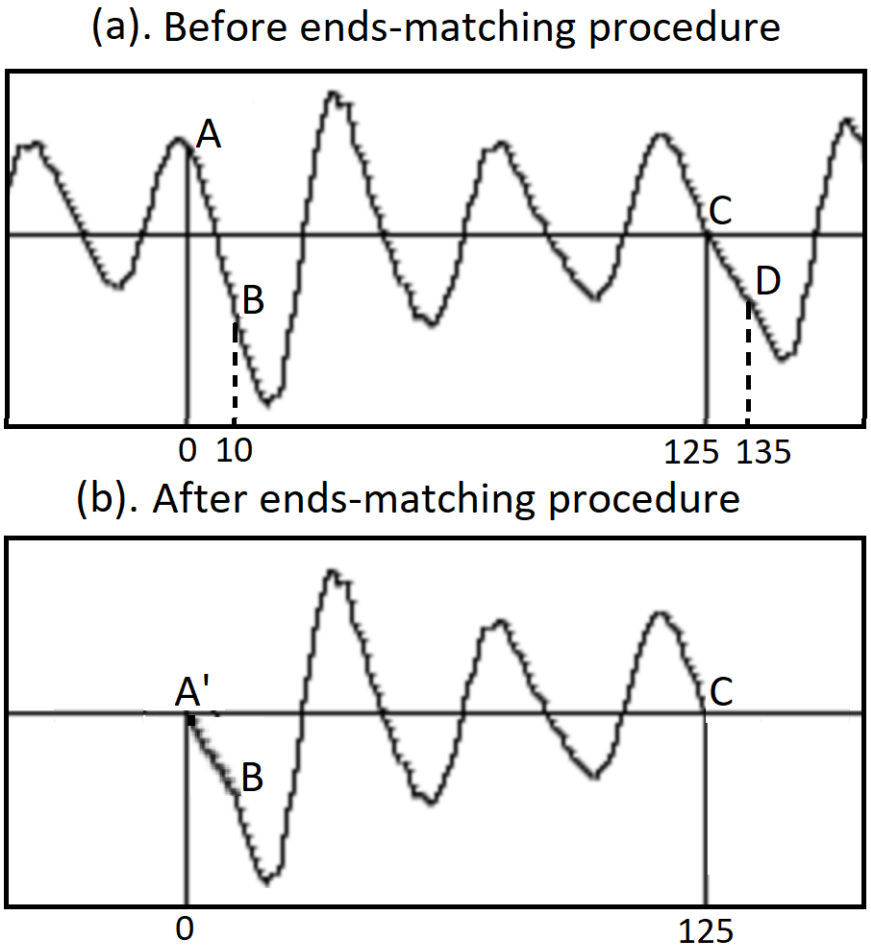

Figure 9. The ends-matching procedure. (a) The starting value A of a pitch period may be not equal to the ending value C. By taking more points after the end, for example 10 points, CD, making a linear combination with the first 10 points AB and then replace the first 10 points with it. (b). The beginning and the end of a pitch period becomes equal, suitable for Fourier analysis. See references[20,21].

A key step is to take a digital Fourier transform on the waveform in each pitch period to generate spectra. In traditional FFT, the length of the data must be a power of 2, such as 256, 512, or 1024. Using the FFT modules in NumPy, a standard package for Python, the length of the array can be any integer. If the input sound wave is an array **wavin** with n points, the code

```
ampspec = np.abs(np.fft.rfft(wavin))
```

creates the amplitude spectrum of **wavin** with $(n//2+1)$ floating point numbers.

For most applications in voice-based diagnostics, vowel sounds are utilized. Here, the biomarkers for all monophthong vowels in US English are extracted. After the notation of ARPABET,[29] there are 10 monophthongs, AA, AE, AH, AO AX, EH, IH, IY, UH, and UW. The diphthongs are represented as transitions from one monophthong to another one. Because the waveform in each individual pitch period contains full information on the timbre of the voice, for each monophthong, the waveform in a single pitch period suffices. Table 1 shows the source of each monophthong, including the sentence number, the time, and the word containing it.

 doi:10.20944/preprints202508.2098.v1

The amplitude spectra of the 10 monophthong vowels are shown in Figures 10 and 12 as blue dotted curves. As shown, those spectra contain rich information. Nevertheless, those amplitude spectra themselves are not appropriate to directly used as biomarkers. As usual, parametrization is necessary. The most used ones are formants and MFCC. In the following two sections, we show how to extract formant parameters and timbre vectors that are pitch-synchronous versions of MFCC.

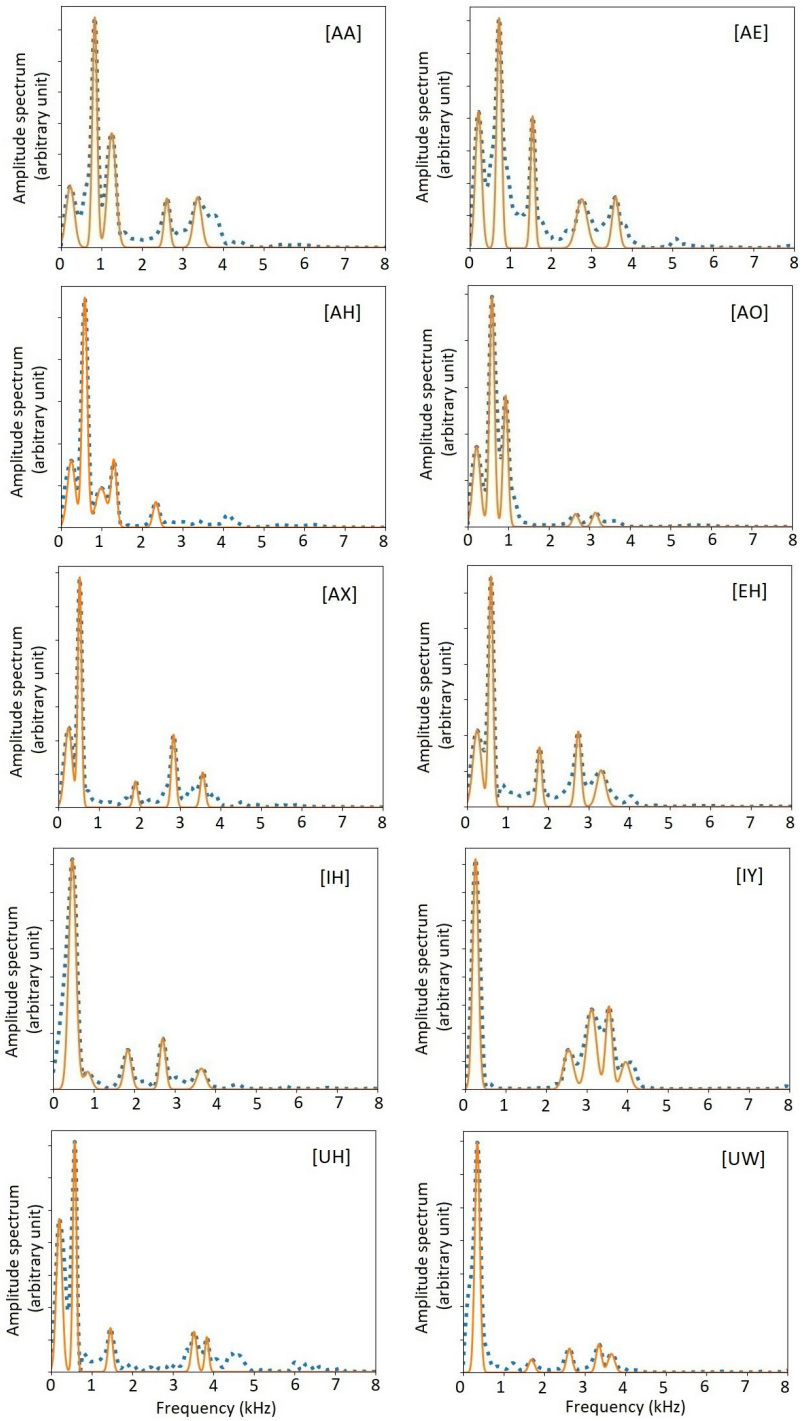

Figure 10. Amplitude spectra and formants of 10 US English monophthong vowels. Blue dotted curves show the original amplitude spectra from individual pitch periods. Red solid curves show the amplitude spectra recovered from the formant data. For clarity, the curves of formants are represented by Gaussian functions.

 doi:10.20944/preprints202508.2098.v1

5. Formant Parameters

Vowels were often characterized by a list of formants. Each formant is characterized by three parameters, central frequency F_n, level L_{Fn}, and bandwidth B_{Fn}.

In the following, a method of extracting formant parameters from the amplitude spectra is described. If a timbron is described by Eq. 2, the amplitude of a complex Fourier transform of each term is

$$\begin{aligned} A_n(f) &= \left| \int_0^\infty C_n e^{-2\pi B_n t - 2\pi i f t} \sin 2\pi F_n t \, dt \right| \\ &= \frac{C_n}{2\pi} \sqrt{\frac{B_n}{B_n^2 + (f - F_n)^2} - \frac{B_n}{B_n^2 + (f + F_n)^2}} \end{aligned} \tag{4}$$

We now pay attention to the values of the amplitude spectrum near $f \cong F_n$. The second term in the square root is much smaller than the first term. Approximately,

$$A_n(F_n) \cong \frac{C_n}{4\pi B_n}. \tag{5}$$

In words, at each peak of the amplitude spectrum, where the frequency is F_n, the amplitude A_n is $C_n/4\pi B_n$. As the first derivative of the curve is zero, the second derivative at the peak position is

$$\left. \frac{d^2 A_n(f)}{df^2} \right|_{f=F_n} \cong \frac{C_n}{4\pi B_n^3}. \tag{6}$$

The second derivative can be estimated from the curve using the finite difference method by a small frequency increment, for example $\delta \cong 50$ Hz,

$$\begin{aligned} V_n &= \frac{1}{\delta^2} [2A_n(F_n) - A_n(F_n + \delta) - A_n(F_n - \delta)] \\ &\approx A_n''(F_n). \end{aligned} \tag{7}$$

Under that approximation, the bandwidth is

$$B_n = \left| \frac{A_n(F_n)}{V_n} \right|^{\frac{1}{2}}. \tag{8}$$

After Eq. 5, the intensity constant C_n is

$$C_n = 4\pi B_n A_n(F_n). \tag{9}$$

Table 1. Source of monophthong vowels

Vowel	Sentence	Time (sec)	Word
AA	a0329	0.36	Ah
AE	b0166	0.44	fast
AH	a0207	0.52	much
AO	b0357	0.45	law
AX	a0388	1.20	ide**a**
EH	a0005	1.25	forg**e**t
IH	a0580	0.29	it
IY	a0329	0.97	deed
UH	a0158	1.16	good
UW	a0102	1.37	soon

Two examples are shown in Table 2. The central frequencies and the bandwidths are in unit of Hz. The amplitude parameter is in PCM unit. The sum of all intensity parameters is 32768. Therefore, while doing voice synthesis, the amplitude of the pcm is in a reasonable range. Intensity values for

 doi:10.20944/preprints202508.2098.v1

formants are very important. For example, by setting the amplitude of F_1 of AA to be 16000, and the amplitude value of F_2 to be small, it will sound like IY, rather than AA.

Table 2. Formants of vowel AA and IY

Vowel AA

Number	F (Hz)	A (pcm)	B (Hz)
F1	251	6028	135
F2	529	17125	66
F3	1918	1714	72
F4	2850	5460	77
F5	3537	2439	119

Vowel IY

Number	F (Hz)	A (pcm)	B (Hz)
F1	258	16389	102
F2	2567	2824	169
F3	3139	5700	167
F4	3571	5897	113
F5	3987	1955	158

Figure 10 shows the amplitude spectra of all US English monophthong vowels, and the spectra recovered from the formants. First, we note that the pitch-synchronous amplitude spectra of those vowels contain rich and detailed information. Second, the amplitude spectra recovered from the formants represent essential information on the amplitude spectra.

To verify the correctness or the accuracy of the formant parameters, a Python program is designed to superpose the timbrons generated by those formant parameters, shown in Eq. 4, to become a melody sung with the vowel by a singer of various voices, including contrabass, bass, tenor, alto, soprano, or child.

Although the origin of the formant parameters is a tenor speaker, for other types of speakers, the formant parameters can be obtained by using a vocal-tract dimension factor κ, see Table 3. Because the formant frequency of a person is inversely proportional to the size of the vocal tract, the central frequency and the bandwidth are divided by the factor κ.

Table 3. Change of speaker identity

Speaker identity	$\overline{f}$(Hz)	MIDI	note	κ
Contrabass	43.7	29	F1	1.35
Bass	61.5	35	B1	1.18
Tenor	123	47	B2	1.0
Alto	175	53	F3	0.85
Soprano	349	65	F4	0.75
Child	494	71	B4	0.68

A list of all formant parameters of the monophthong vowels is included in an appendix of this paper. To test the accuracy of the formant parameters by singing synthesis, a Python program

`melody.py`

is included. For example, using the command

```
python3 melody.py Grieg AA 67 S 60
```

a wav file of the first phrase of Eduard Grieg's Morning Mode in Peer Gynt by a soprano singer in vowel AA on G major with tempo 60 beats per minute is created.

 doi:10.20944/preprints202508.2098.v1

6. Timbre Vector and Timbre Distance

The formant parameters are intuitive and can be directly used for voice synthesis. However, it is not the best for voice-based diagnostics. The numbering of formants has a certain arbitrariness. For example, the strongest formant of vowel [AA] is at about 730 Hz.[10] It is typically identified as the first formant. But, there is often a weak formant at about 270 Hz. There is no objective criterion to label the 270 Hz peak as the first formant or to label the 730 Hz peak as the first formant. Furthermore, the method represented by Eqs. 5 through 9 is based on an independent formant model, which requires that the frequency distance of adjacent formants is greater than the bandwidth of the formants. Otherwise, two formant peaks may merge. Some formants cannot be identified by the peaks in the amplitude spectrum.

To find a better mathematical representation of the amplitude spectrum, attention is directed to the experience of voice-based diagnostics: among all pitch-asynchronous biomarkers, MFCC is the best, due to the implementation of the Mel scale.[1–6] The pitch-synchronous equivalent of MFCC is *timbre vectors*, by implementing the Mel scale using a standard mathematical tool of Laguerre functions,[32] see Chapter 6 of *Elements of Human Voice*.[20] Here is a brief presentation.

6.1. Laguerre Functions

A Laguerre function is the product of a weighing function with a Laguerre polynomial.[32] The first two Laguerre polynomials are

$$L_0(x) = 1, \tag{10}$$

$$L_1(x) = 1 - x. \tag{11}$$

For $n \geq 2$, the Laguerre polynomials are defined by the recurrence relation[32]

$$L_n(x) = \frac{2n-1-x}{n} L_{n-1}(x) - \frac{n-1}{n} L_{n-2}(x). \tag{12}$$

Using a simple program, all Laguerre polynomials can be generated. A Laguerre functions is the product of a weighing function and a Laguerre polynomial

$$\Phi_n(x) = e^{-x/2} L_n(x). \tag{13}$$

They are orthonormal on the interval $(0, \infty)$,

$$\int_0^\infty \Phi_m(x)\, \Phi_n(x)\, dx = \delta_{m,n}. \tag{14}$$

The waveforms are shown in Figure 11. As shown, the behavior of the Laguerre functions closely resemble the behavior of the Mel scale.[32] The first few Laguerre functions describe the gross features of the low-frequency amplitude spectrum. The later Laguerre functions represent the high-frequency spectrum and refine the low-frequency spectrum. Overall, the frequency resolution in the range of 0 kHz to 1 kHz is high, and the resolution at higher frequencies gradually becomes lower.

 doi:10.20944/preprints202508.2098.v1

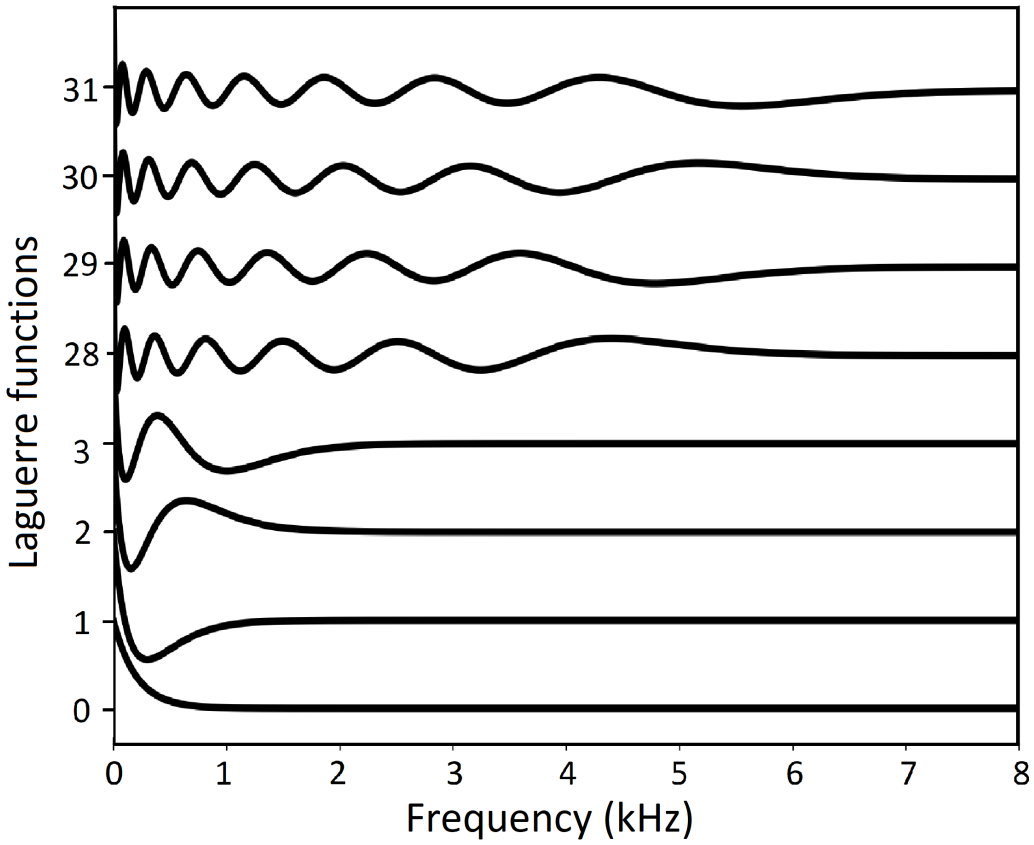

Figure 11. Laguerre functions. The analytic version of the Mel frequency scale. The first few Laguerre functions describe the gross features of the low-frequency amplitude spectrum. The later Laguerre functions represent the high-frequency spectrum and refine the low-frequency spectrum. The mathematical form of the Laguerre functions, Eqs. 10 through 13, is simpler than the definition of MFCC.[32]

 doi:10.20944/preprints202508.2098.v1

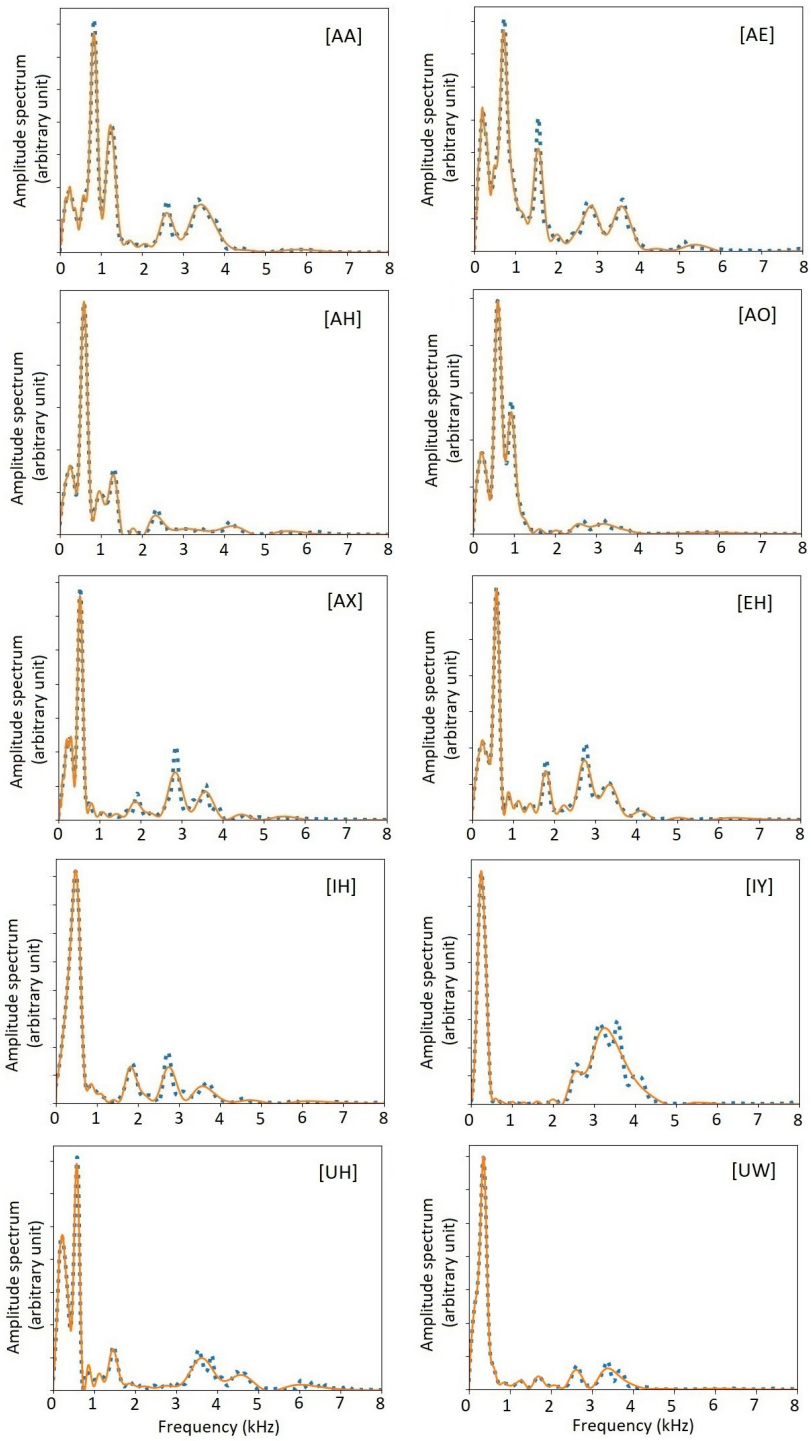

Figure 12. Amplitude spectra of 10 US English monophthong vowels. Blue dotted curves show the original amplitude spectra from individual pitch periods. Red solid curves show the amplitude spectra recovered from a timbre-vector expansion. Following the Mel curve of human perception, the resolution at lower frequencies is finer than at higher frequencies.

6.2. Definition of Timbre Vector

The amplitude spectrum of a timbron $A(f)$ can be approximated by a sum of N such Laguerre functions,

$$A(f) \cong \sum_{n=0}^{n<N} C_n \Phi_n(\kappa f), \tag{15}$$

 doi:10.20944/preprints202508.2098.v1

where the coefficients C_n are

$$C_n = \int_0^\infty \kappa A(f)\, \Phi_n(\kappa f)\, df. \tag{16}$$

The scaling factor κ is a constant with a dimension of time, having a convenient unit of msec, chosen to optimize accuracy. The final results are pretty insensitive to its value. For example, ranging from κ = 0.1 msec to κ = 0.3 msec, the final results are virtually identical.

The coefficients C_n form a vector, denoted as **C**. The norm of the vector

$$\| \mathbf{C} \| = \sqrt{\sum_{n=0}^{n<N} C_n^2} \tag{17}$$

represents the overall amplitude. The *normalized Laguerre spectral coefficients*, constitute a *timbre vector*

$$\mathbf{c} = \frac{\mathbf{C}}{\| \mathbf{C} \|}, \tag{18}$$

represents the spectral distribution of the pitch period, characterizing the timbre of that pitch period independent of pitch frequency and intensity. Obviously, the timbre vector is normalized:

$$\sum_{n=0}^{n<N} c_n^2 = 1. \tag{19}$$

In this paper, we use 32 Laguerre functions, to generate a 32-componant timbre vector. To unify and save storage space, each component of the timbre vector is multiply by 32768 and expressed as type numpy.int16 in Python. Because the pcm is also in numpy.int16, it is sufficiently accurate. Each timbre vector is then stored as 64 bytes of binary data. To compute timbre distance, as shown later, the stored values are first divided by 32768 to become floating-point numbers.

Figure 12 shows the amplitude spectra of the US English monophthongs and the spectra recovered from the timbre vectors. As shown, the agreement is excellent, showing the accuracy of the timbre-vector representation. As biomarkers, the timbre vectors have significant advantages over the formants. The numbering of formants is somewhat arbitrary. It is difficult to compare two sets of formants. The timbre vectors are generated automatically and objective. Especially, there is a well-defined and well-behaved timbre distance between two timbre vectors from two pitch periods, as shown below.

6.3. Timbre Distance

A critical parameter in any parametrical representation of speech signal is the distortion measure, or simply distance. Because the timbre vectors are independent of pitch frequency and intensity, the distances defined here can be justly characterized as *timbre distance* between a pair of timbre vectors **x** and **y**.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{n=0}^{n<N} (x_n - y_n)^2}, \tag{20}$$

where x_n and y_n are the components of the two timbre vectors, or the arrays of the normalized Laguerre spectral coefficients **x** and **y**, see Eq. 15 through 18. The accurate definition of timbre distance makes the timbre vector a prime choice for voice-based diagnostics.

A list of timbre distances among the 10 US English monophthong vowels is shown in Table 4. As expected, the values differ significantly for various pairs. The maximum timbre distance is 2, when all components of one vowel is non-zero, the correspondent timbre vector elements of another vowel is zero. As shown in Table 4, many timbre distances of pairs of vowels are greater than 1, indicating a significant spectral difference. The timbre distance between [IH] and [AX] is small. This is expected because those two phones are often interchangeable.

 doi:10.20944/preprints202508.2098.v1

Table 4. Timbre distances among US English monophthong vowels

	AA	AE	AH	AO	AX	EH	IH	IY	UH	UW
AA	0.000	0.453	0.939	0.719	1.219	1.032	1.246	1.214	1.102	1.297
AE	0.453	0.000	0.631	0.411	0.806	0.579	0.810	0.938	0.617	0.948
AH	0.939	0.631	0.000	0.183	0.436	0.211	0.515	1.335	0.269	1.106
AO	0.719	0.411	0.183	0.000	0.654	0.334	0.698	1.339	0.445	1.178
AX	1.219	0.806	0.436	0.654	0.000	0.305	0.191	0.962	0.290	0.861
EH	1.032	0.579	0.211	0.334	0.305	0.000	0.397	1.002	0.271	0.943
IH	1.246	0.810	0.515	0.698	0.191	0.397	0.000	0.896	0.354	0.436
IY	1.214	0.938	1.335	1.339	0.962	1.002	0.896	0.000	0.747	0.381
UH	1.102	0.617	0.269	0.445	0.290	0.271	0.354	0.747	0.000	0.653
UW	1.297	0.948	1.106	1.178	0.861	0.943	0.436	0.381	0.653	0.000

In Section V, we have shown how to do voice synthesis using formants. Using timbre vectors, voice synthesis can also be implemented, and it is often even better than using formants. Details are shown in Chapter 7 of the reference book[20]. Using dispersion relations, the phase spectrum can be recovered from the amplitude spectrum. Using reverse FFT from Numpy, individual timbrons can be recovered. By superposing the timbrons according to a music score, singing can be synthesized.

7. Jitter, Shimmer, and Spectral irregularity

Jitter and shimmer are often used as biomarkers for voice disorders,[33,34] and now also as biomarkers for other medical conditions, see a recent review.[35] Jitter means the frequency variation of the sound wave over adjacent pitch periods, and shimmer means the amplitude variation of the sound wave over adjacent pitch periods. Within the framework of the source-filter theory of voice production, there are no valid definitions of jitter and shimmer. Both jitter and shimmer are described as perturbations to the source of voice, which originally should have a constant pitch.[34,35]

Usually, jitter and shimmer are measured by asking the patient to produce several vowels in a steady frequency and volume, then by looking on the voice waveforms using a dedicated software, such as SEARP,[34] or a free software such as PRAAT.[36] For certain vowels, in each pitch period, there is a unique sharp peak at the same time instant in each pitch period, see Figure 13(a). The time distance between two adjacent peaks is defined as the pitch period. And the intensity of the period is determined by the value of the peak. However, for some vowels such as [IY] and [EH], the waveform has multiple peaks, and the determination of the pitch period is compromised, see Figure 13(b). On the other hand, for other vowels, for example [OH] and [UW], there are no sharp peaks in the voice waveform, and the determination of the pitch period is also compromised, see Figure13(c).

According to the timbron theory of voice production, human voice is generated a pitch period at a time. Each glottal closure instant triggers an decaying elementary acoustic wave called a timbron. Because the timing and the intensity of each timbron is independent, jitter and shimmer are intrinsic to the nature of human voice. Based on the timbron theory, the methods of measuring jitter and shimmer can be precisely defined in terms of pitch-synchronous analysis method, see Figure 13(a), (b), and (c), based on the GCI in each pitch period. If simultaneously acquired EGG signals are available, such as the Saarbrucken voice database,[28] GCIs can be accurately determined. Such measurements are objective and reproducible, important for diagnostics.

In the following, we study in detail three recordings in the Saarbrucken database, the vowels [a], [i], amd [u] spoken by speaker 88, see Figure 14. There are simultaneously acquired EGG signals. The derivative of the EGG signal over time is shown in Figure 14(a). As shown, in each pitch period, there is a prominent peak, especially a sharp rising edge, signifies the beginning of a pitch period. As the sampling rate of the Saarbrucken database is 50 kHz, the beginning of a pitch period, or the GCI, can be determined to an accuracy of 0.02 msec. See the red dashes in Figure 14(b). In Figure 14(b), it is shown that the waveform of each pitch period is similar. By taking a Fourier transform to the waveform in each pitch period, an amplitude spectrum is obtained, as shown in Figure 14(c) as the dotted blue curve. For each pitch period, a timbre vector is obtained using the Laguerre function expansion. Then, the amplitude spectrum is recovered using Eq. 15. Shown by solid red curves, the recovered spectrum

 doi:10.20944/preprints202508.2098.v1

matches perfectly with the original amplitude spectrum. Note that in each recording, the number of independent pitch periods is large. As shown in Table 5 and Figure 14, the number of timbre vectors are 308, 277, and 172 for the vowels [a], [i], and [u], respectively. The number is much greater than the number of MFCCs, and each one has a much higher resolution.

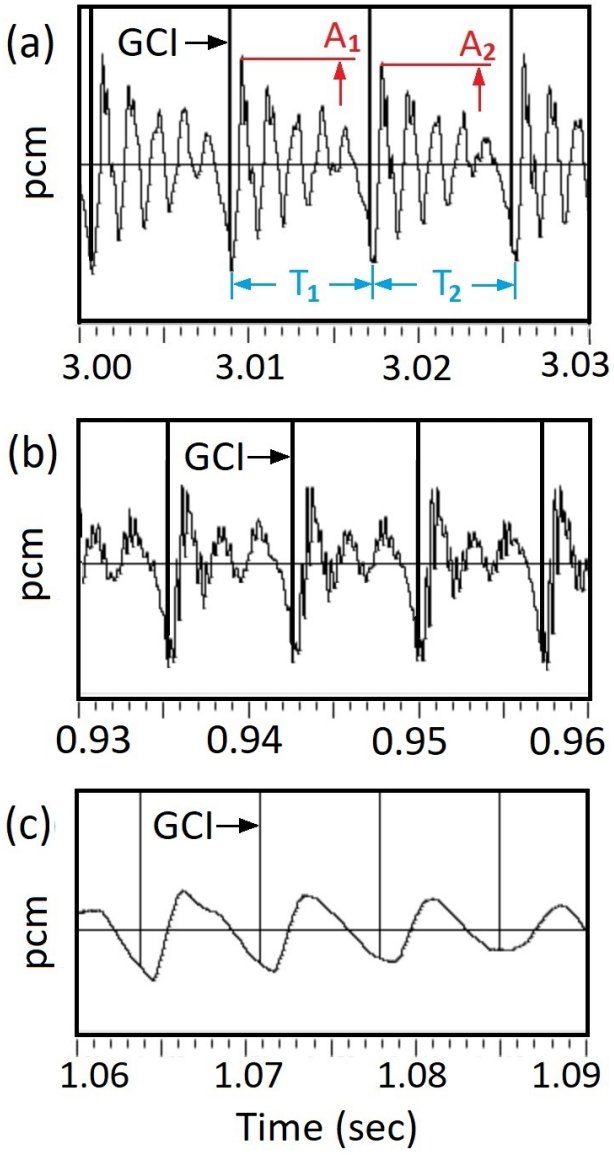

Figure 13. Measurement of jitter and shimmer Typical method of measuring jitter and shimmer is to investigate the peaks in the voice waveforms by a dedicated software, which is not reliable. It is better to rely on the GCIs, either derived from EGG signals or detected from the sound wave.

 doi:10.20944/preprints202508.2098.v1

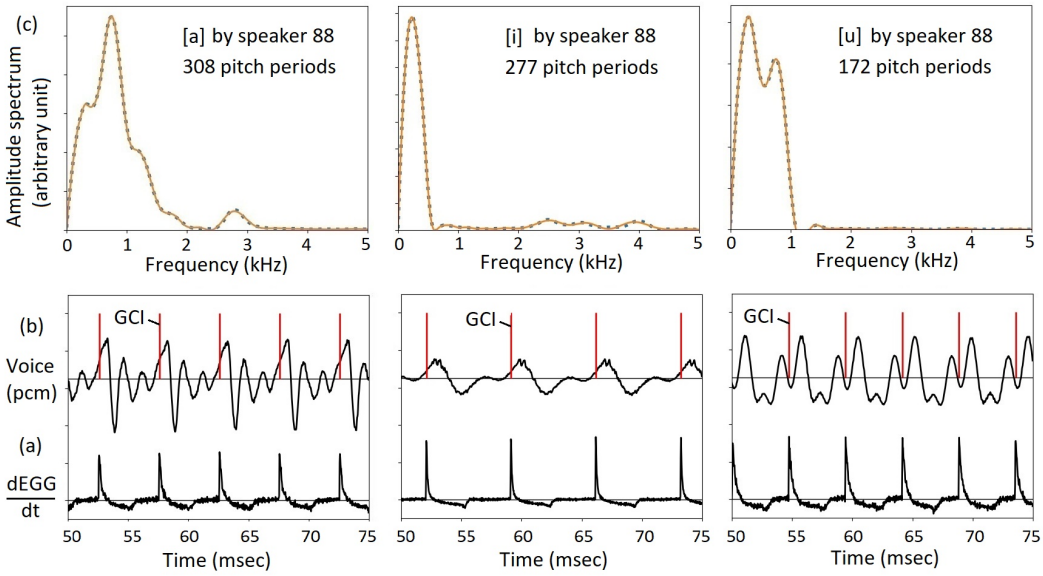

Figure 14. Vowel recordings in the Saarbrucken database. (a) the derivatives of EGG, showing sharp peaks at the GCIs. (b) waveforms, a nearly perfect repetition over all pitch periods. (c) amplitude spectra in blue dotted curves, and that recovered from the timbre vectors in red curves. For each vowel, the amplitude spectra in all pitch periods are very similar.

Following the timbron theory of voice production, the pitch period is the time difference between two adjacent GCIs, a series of pitch periods are found for each vowel. Denoting them as T_i, $i = 1, 2, 3, ...N$, where N is the number of pitch periods, following the usual definition of relative jitter,[34,35] we have

$$\text{jitt} = \left[\frac{1}{N-1}\sum_{i=1}^{N-1}|T_{i+1} - T_i|\right]\left[\frac{1}{N}\sum_{i=1}^{N}T_i\right]^{-1}. \tag{21}$$

Usually it is expressed as percentages. For the three recordings from speaker 88 in the Saarbrucken voice database,[28] the number of jitter are shown in Table 5.

For the shimmer, following the standard definition of decibels, the energy level in each pitch period can be defined by adding the energy of each pcm point,

$$L_i = 10\log_{10}\sum_{m=1}^{M_i} p_m^2 \tag{22}$$

where p_m is the m-th pcm in period i having a total number M_i of pcm points. The unit of pcm will cancel out. The definition is independent of the polarity of pcm. The shimmer in terms of decibel is

$$\text{ShdB} = \frac{1}{N-1}\sum_{i=1}^{N-1}|L_{i+1} - L_i|. \tag{23}$$

For the three vowel recordings from speaker 88 in the Saarbrucken voice database,[28] the numbers of shimmer are shown in Table 5.

Table 5. Jitter, shimmer, and spectral irregularity based on the GCI values derived from EGG signals

Vowel	Number of periods	Jitter percent	Shimmer dB	Spectral irregularity
a	308	0.652	0.197	0.376
i	277	0.401	0.089	0.070
u	172	0.576	0.235	0.250

 doi:10.20944/preprints202508.2098.v1

There is yet another parameter that can be precisely defined, the *spectral irregularity*. The timbre in each pitch period can be precisely represented by its timbre vector, which is *normalized* with a well-defined timbre distance, see Eqs. 19 and 20. The timbre distance between two adjacent pitch periods

$$d_{i,i-1} = \sum_{n=1}^{32} \left[c_n^{(i+1)} - c_n^{(i)} \right]^2 \tag{24}$$

is always non-negative. Therefore, a simple average can be applied. The average spectral irregularity of a voice recording with N pitch periods can be defined as

$$\text{Timbvar} = \frac{1}{N-1} \sum_{i=1}^{N-1} \sum_{n=1}^{32} \left[c_n^{(i)} - c_n^{(i-1)} \right]^2. \tag{25}$$

For the three recordings from speaker 88 in the Saarbrucken voice database,[28] the spectral irregularities are shown in Table 5. The existence of well-defined timbre distances among all pitch periods is a great advantage of timbre vectors.

8. Detecting GCI from Voice Signals

For voice recordings with simultaneous acquired EGG signals, for example the CMU ARCTIC database[27] and the Saarbrucken database,[28] GCIs can be easily identified to implement pitch synchronous analysis. Many voice recordings do not have simultaneous acquired EGG signals. Nevertheless, because of the importance of pitch synchronous speech signal processing, since the 1970s, methods to detect GCIs from voice signals have been developed. A recent review evaluated and compared five most successful methods to decent GCIs from voice signals based on the analysis of voice waveforms.[37] By applying those methods to the CMU ARCTIC database and taking the EGG based GCI as the reference, it is found that many methods can produce more than 80% of GCIs within 0.25 msec from the reference GCIs.[37] It is often sufficiently accurate for voice-based diagnostics.

In recent years, methods based on AI, using conventional neural network (CNN) are applied.[38] The results are compared with the CMU ARCTIC database, and the percentage of GCIs within ± 0.25 msec from the EGG references often exceeds 95%.

Here we present a simple and reliable method to detect GCIs from voice signals.[20] It can be implemented as a Python code of less than one page. It uses an asymmetric window function $w(n)$ to find the peak amplitudes in each pitch period to generate the GCIs,

$$w(n) = \sin \frac{\pi n}{N} \left(1 + \cos \frac{\pi n}{N} \right), \qquad -N < n < N, \tag{26}$$

where N is the width of the asymmetric window, see Figure 15. At the ends, $n = -N$ and $n = N$, the asymmetric window function approaches zero as the third power of the distance to the ends, which makes it smooth.

 doi:10.20944/preprints202508.2098.v1

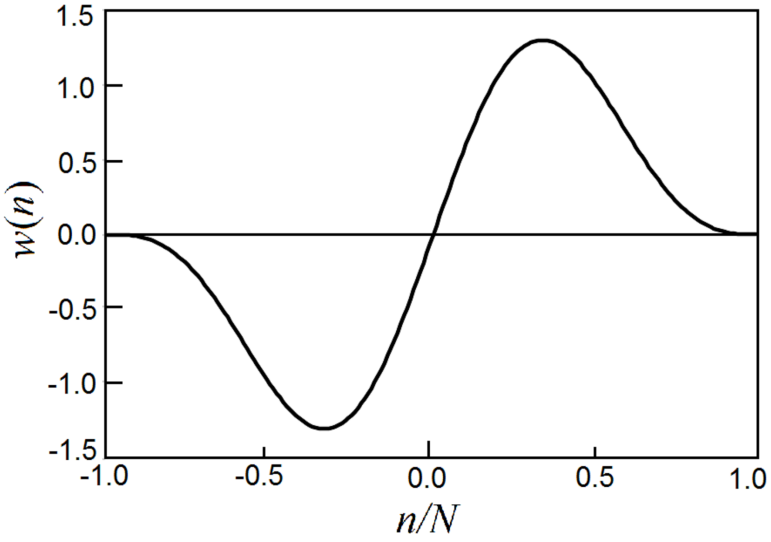

Figure 15. Asymmetric window for detecting GCI. By convoluting the derivative of the voice signal with an asymmetric window, a profile function is generated. The peaks in the profile function point to GCIs. See Figure 16.

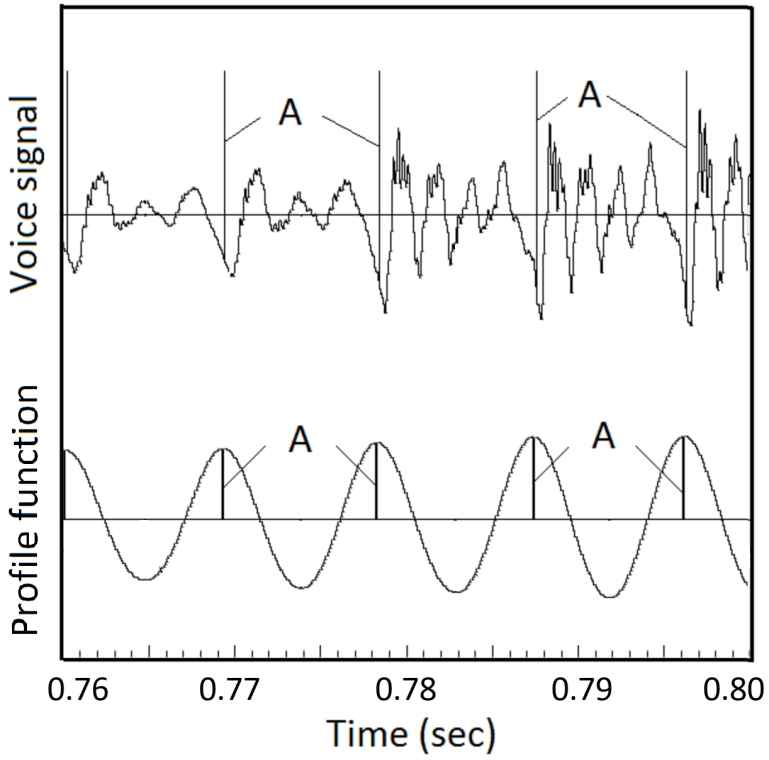

Figure 16. Profile function for detecting GCIs from the sound wave. An example of the profile function generated by convoluting the derivative of the voice signal with an asymmetric window, shown in Figure 15. The peaks of the profile function, A, are identified as GCIs.

To find segmentation points, the derivative of the voice signal is convoluted with the asymmetric window to generate a profile function, see Figure 16. The peaks of the profile function point to GCIs. By applying this method on the three Saarbrucken voice recordings, the GCI points from the voice signals are consistently within ± 0.2 msec from the GCIs derived from EGG signals.

The width of the asymmetric window N is an input parameter. However, the results are fairly insensitive to the value of N. For example, for the vowels [a] and [u] by speaker 88 of the Saarbrucken database, using N between 130 and 330 (3.1 msec to 6.6 msec), the GCIs produced are virtually identical. Therefore, to execute that program, it suffices to have an estimate of the average pitch of the voice

 doi:10.20944/preprints202508.2098.v1

recording to an accuracy of -30% to +50%. This can be done by using open source software such as Praat, a smart phone app such as Cleartune, writing a dedicated program or a small AI module. As a rule of thumb, using the average pitch period, 8 msec for men and 4 msec for women as the width, the GCIs produced are reliable except for extreme cases. Because the voice samples used in dignostics typically have a small range of pitch variation, the method of detecting GCIs based on an asymmetric window should work.

By using the GCI points detected from voice signals as segmentation points, the values of jitter, shimmer, and spectral irregularity are reevaluated. The results are similar, see Table 6. By comparing with Table 5, one finds that the values of jitter are smaller. The values of shimmer and spectral irregularity are virtually identical to those obtained from the EGG data. Therefore, the GCIs detected from voice signals are basically reliable for extracting pitch-synchronous biomarkers.

On the other hand, the EGG signals are still valuable for diagnostics. First, although timbre vectors, shimmer, and spectral irregularities measured based on the GCIs detected from voice waveforms are virtually identical to those from EGG signals, some symptoms related to jitter could be missed. The pitch periods measured from EGG signals are more accurate. Second, the EGG signals contain additional valuable information not available from voice signals, including glottal opening instants and pathological glottal behaviors, such as multiple closings in a pitch period and incomplete glottal closures. Therefore, if possible, to record voice and EGG signals simultaneously is a better strategy for diagnostics.

Table 6. Jitter, shimmer, and spectral irregularity based on the GCI values derived from voice signals

Vowel	Number of periods	Jitter percent	Shimmer dB	Spectral irregularity
a	308	0.320	0.194	0.324
i	277	0.258	0.083	0.088
u	172	0.475	0.220	0.396

9. Results and Discussions

Voice analysis combined with AI is rapidly becoming a vital tool for disease diagnosis and monitoring. A key issue is the identification of vocal biomarkers, that are quantifiable features extracted from the voice to assess a person's health status or predict the likelihood of certain diseases. Currently, the biomarkers are extracted using pitch-asynchronized methods that need improvement. Based on the timbron theory of voice production, a pitch-synchronous method of vocal biomarker extraction is presented, including the measurement of all formant parameters and the timbre vectors, as well as jitter, shimmer, and spectral irregularities. The biomarkers extracted using pitch-synchronous analysis contain abundant, accurate, objective, and reproducible information from the voice signals that can improve the usability and reliability of voice-based diagnostics.

Supplementary Materials: More details of the theory and methods presented in this paper are contained in Reference [20] and documented in US Patents 8,719,030, 8,744,854, and 9,135,923. For access to the source codes written in Python and data files related to the current article and those publications please contact Columbia Technology Venture.

Funding: This research received no external funding.

Data Availability Statement: All data used here is publicly available.

Conflicts of Interest: The author declares no conflicts of interest.

 doi:10.20944/preprints202508.2098.v1

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
MFCC	Mel-frequency cepstral coefficients
LFCC	Linear frequency cepstral coefficients
LPC	Linear prediction coefficients
EGG	Electroglottograph
GCI	Glottal closing instant
FFT	Fast Fourier transform

References

1. Sankaran A, Kumar L, Advances in Automated Voice Pathology Detection: A Comprehensive Review of Speech Signal Analysis Techniques. IEEE Access, 12, 181127 (2024).
2. Fagherazzi G, Fischer A, Ismael M, and Despotovic V, Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice, Biomarkers from Research to Clinical Practice, Digital Biomarkers 5, 78-88 (2021).
3. De Silva U, Madanian S, Olsen S, Templeton J, Poellabauer C et al, Clinical Decision Support Using Speech Signal Analysis: Systematic Scoping Review of Neurological Disorders, Journal of Medical Internet Research, 27, e63004 (2025).
4. Kapetanidis P, Kalioras F, Tsakonas C, Tzamalis P, Kontogiannis G, and Stavropoulus T. et al, Respiratory Diseases Disgnosis Using Audio Analysis and Artificial Intelligence: A Systematic Review. Sensors, 24, 1173 (2024).
5. Ankishan H, Ulucanlar H, Akturk I, Kavak K, Bagci U, and Yenigun B, Early Stage Lung Cancer detection from Speech Sounds in Natural Environments, https://doi.org/10.1016/j.bspc.2024.106628.
6. Bauser M, Kraus F, Koehler F, Rak K, Pryss R, Weiss C, et al. Voice Assessment and Vocal Biomarkers in Heart Failure: A Systematic Review. Circulation: Heart Failure, August 2025.
7. Mekyska J, Janousova E, Gomez-Vilda P, Smekal Z, Rektorova I, Eliasova I, and Mrackova M, Robust and Complex Approach of Pathological Speech Signal Analysis, Neurocomputing, 167: 94-111 (2015).
8. Baghai-Ravary L and Beet SW, Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders, Springer 2013.
9. Wszolek W, Selected Methods of Pathological Speech Signal Analysis, Archives in Acoustics, 31:413-130 (2006).
10. Rabiner LR and Schafer RW, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliff, New Jersey 1978).
11. Rabiner LR and Juang BH, *Fundamentals of Speech Recognition* (Prentice-Hall, New Jersey 1993).
12. Fant G, *Acoustic Theory of Speech Production* (Molton & Co., The Hagues, The Netherlands 1960).
13. Daly I, Hajaiej Z, and Gharsallah A, Physiology of Speech/Voice Production, Journal of Pharmaceutical Research International, 23:1-7 (2018).
14. Davis SB and Mermelstein P, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences. IEEE Transaction on ASSP. 28:357-366 (1980).
15. Nagraj P, Jhawar G, and Mahalakshmi KL, Cepstral Analysis of American Vowel Phonemes (A Study on Feature Extraction and Comparison), IEEE WISPNET Conference (2016).
16. Hess WJ. A pitch-synchronous digital feature extraction system for phonemic recognition of speech. *IEEE Transactions on ASSP*, ASSP-24:14–25, 1976.
17. Medan Y and Yair E. Pitch synchronous spectral analysis scheme for voiced speech. *IEEE Trans. ASSP*, 37:1321–1328, 1989.
18. Svec JG, Schutte HK, Chen CJ, and Titze IR, "Intgegrative insights into the myoelastic theory and acoustics of phonation. Scientific tribute to Donald G. Miller", Journal of Voice, 37:305-313 (2023).
19. Chen CJ and Miller DG, "Pitch-Synchronous Analysis of Human Voice", Journal of Voice, 34:494-502 (2019).
20. Chen CJ, *Elements of Human Voice* (World Scientific Publishing 2016).
21. Aaen M, Mihaljevic I, Johnson AM, Howell I, and Chen CJ, A Pitch-Synchronous Study of Formants, Journal of Voice (April 2025).
22. Baken, RJ "Electroglottography." Journal of Voice, 6: 98-110 (1992).
23. Herbst CT, "Electroglottography - An Update." Journal of Voice, **34** 503-526 (2020).

 doi:10.20944/preprints202508.2098.v1

24. Miller DG, “Resonance in Singing", Inside View Press (2008).
25. Fabre P, Un procédé électrique percutané d’inscription de l’acclement glottique au cours de la phonetion: glottographie de haute fréquence. premiers résultats. *Bulletin de l’Académie nationale de médecine*, 141:66–69, 1957.
26. Fabre P. La glottographie électrique en haute fréquence, particuralités de l’appareillage. *Comptes Rendus, Sociéte de Biologie*, 153:1361–1364, 1959.
27. Kominek J and Black A. CMU ARCTIC Databases for Speech Synthesis. *CMU Language Technologies Institute, Tech report*, CMU-LTI-03-177, 2003.
28. Woldert-Jokisz B, Saarbrucken voice database, Inst. Phonetics Saarland Univ. Saarbrucken, Germany, Tech. Rep. 2007.
29. See Wikipedia entry of ARPABET. In some publications, the r-colored phones ER and AXR are also listed as monophthongs. In CMU pronunciation dictionaries, there is no AXR.
30. Atal BS and Hanauer LS, “Speech Analysis and Synthesis by Linear Prediction of the Speech Wave,” J. Acoust. Soc. Am. 50:637-655 (1971).
31. Markel JD and Gray AH, *Linear Prediction of Speech* (Springer Verlag, New York 1976).
32. Arfken G, *Mathematical Methods for Physicists.* Academic Press, New York, 1966.
33. Brockmann M, Drinnan MJ, Storck C, and Carding PN. Reliable Jitter and Shimmer Measurements in Voice Clinics: The Relevance of Vowel, Gender, Vocal Intensity, and Fundamental Frequency Effects in a Typical Clinical Task. Journal of Voice, 25:(1) 44-53 (2011).
34. Teixeria JP, Oliveria C, Lopes C, Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters, Procedia Technology, 9:1112-122 (2013,).
35. Jurca L and Viragu C, Jitter and Shimmer Parameters in the Identification of Vocal Tract Pathologies. Current State of Research. In Springer Proceedings in Physics, Acoustics and Vibration of Mechanical Structures – AVMS-2023, pp 155-164 Springer 2025.
36. Boersma P and van Heuven V, Speak and unSpeak with Praat, Glot International, 5:341-347 (2001).
37. Drugman T, Thomas M, Gudnason J, Naylor P, and Dutoit T, Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review. arXiv:2001.00473v1.
38. Yang S, Wu Z, Shen B, and Meng H, Detection of Glottal Closure Instants from Speech Signals: A Convolutional Neural Network Based Method. Proceedings of Interspeech 2018, 317-321.