Pitch-Synchronous Analysis of Human Voice

*C. Julian Chen, and †Donald A. Miller, *New York, New York, and †Groningan, The Netherlands

Summary: Objective. Based on simultaneous voice and electroglottograph (EGG) signals, to gain a better understanding of human voice production process, to make pitch-synchronous segmentation of voice signals, and to make visual representations of pitch marks and timbre spectra with high resolution.

Methods/Design. The traditional spectrogram segments the voice signals with a process window of fixed size and fixed shift, then performs fast Fourier transformation after multiplied with a window function, typically a Hamming window. Then display power spectrum in both frequency and time. Pitch information and timbre information are mixed. The new design segments the signals into pitch periods, either using the derivatives of the EGG signals or based on the voice signals, then performs Fourier analysis to the segment of signals in each pitch period without using a window function. The pitch information and the timbre information are cleanly separated. The graphical representations of both pitch marks and timbre spectra exhibit high resolution and high accuracy.

Results. Detailed analysis of simultaneously acquired voice and EGG signals provides a more precise understanding of human-voice production process. The transient theory of voice production, proposed by Leonhard Euler in early 18th century, is substantiated with modern data. Based on the transient theory of voice production, a pitch-synchronous spectrogram software is developed, which makes a visual representation of pitch marks and timbre spectra. In addition, the timbre spectrum and the power evolution pattern in each pitch period can be displayed individually.

Conclusions. Simultaneously acquired voice and EGG signals indicates that each glottal closing triggers a decaying elementary wave in the vocal tract. A superposition of those elementary waves constitutes voice. Based on that concept and using EGG data, a pitch-synchronous voice signal processing method is developed. The voice signal is first segmented into pitch periods, then the two ends are equalized. Fourier analysis is applied to obtain the timbre spectrum of each pitch period. High resolution display of timbre spectrum is generated. The power evolution pattern in each pitch period is also displayed.

Key Words: Human voice−Production−Analysis−Pitch period−Timbre spectra−Graphical display.

TRANSIENT THEORY AND STEADY-STATE THEORY OF HUMAN VOICE

Two theories of human voice production, the transient theory and the steady-state theory, coexisted for more than 150 years. Descriptions of the theories can be found in a 2011 monograph *Speech Spectrum Analysis* by S. A. Fulop.[1] In many classical monographs, including T. Chiba and M. Kajiyama's *The Vowel, Its Nature and Structure*,[2] E. W. Scripture's *The Elements of Experimental Phonetics*,[3] P. Ladefoged's *Elements of Acoustic Phonetics*,[4] and Lord Rayleigh's magnum opus *Theory of Sound*,[5] the two theories are described and compared in detail.

The transient theory of human voice production was proposed by Leonhard Euler[6] in 1727. To start, the vocal folds emit an impulse. The impulse triggers a transient response in the vocal tract to produce a decaying wave. A series of impulses produce a series of decaying waves. The superposition of those decaying waves makes voiced sound. Pitch frequency is the repetition rate of impulses. Euler's theory was experimentally verified by R. Willis[7] in 1831. After Thomas Edison invented the phonograph, German physiologist L. Hermann studied the recorded waveforms of human voice in detail, further validated the transient theory.[1] Hermann coined the term "formant" in his papers, then used by voice scientists ever since.[1] His results are summarized by E. W. Scripture in early 20th century.[3]

The steady-state theory was proposed in 1837 by C. Wheatstone in a comment on Willis's theory.[8] While agreed in every point with Willis, Wheatstone proposed a simpler version of voice production theory. If the repetition rate of the impulses is a constant, a fundamental frequency can be defined. Since the voice signal is then strictly periodic, Fourier analysis can be applied to compute the overtones. Timbre can be expressed as the intensity envelop of overtones. In 1865, H. Helmholtz[9] systematically presented the steady-state theory. After the availability of digital computers in 1960s, a version of the steady-state theory, the source-filter theory, was published by G. Fant,[10] which enables a straightforward digital processing of speech signals on computers, as shown below.

PITCH-ASYNCHRONOUS ANALYSIS OF HUMAN VOICE

To date, a typical digital processing of voice signals is as follows.[11] First, the voice signal is segmented into overlapping frames with a fixed length (typically 25 msec) and a fixed shift (typically 10 msec). Each segment is multiplied by a

Accepted for publication January 16, 2019.
Presented at the 2018 Annual Symposium of The Voice Foundation May 31, 2018, The Westin, Philadelphia, PA
From the *Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York; and the †Groningen Voice Research, Groningan, The Netherlands.
Address correspondence and reprint requests to C. Julian Chen, Columbia University, New York, NY, 10027. E-mail: jcc2161@columbia.edu
Journal of Voice, Vol. 34, No. 4, pp. 494−502
0892-1997

https://doi.org/10.1016/j.jvoice.2019.01.009

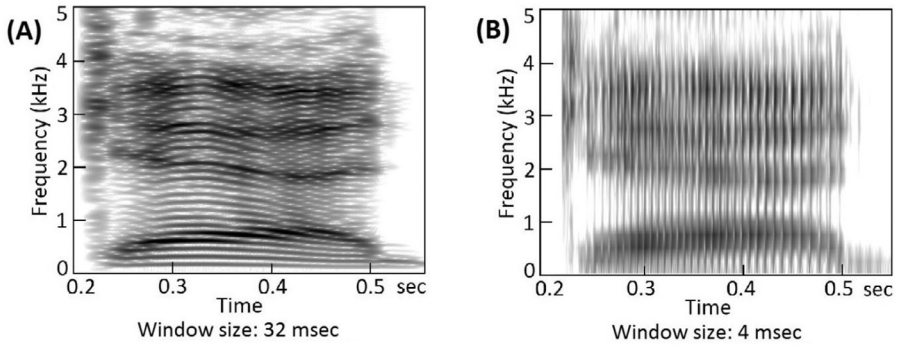

FIGURE 1. Dependence of spectrogram features with process window size.

window function, typically a Hamming window. Then a fast Fourier transformation (FFT) is applied. The output is the FFT amplitudes at frequencies of multiples of the inverse of the window size. Pitch information and timbre information are mixed.

Based on the above pitch-asynchronous digital processing, a widely used visual representation, the spectrogram, was developed. It is an essential component of many voice-and-speech signal analysis software packages, for example, Wavesurfer,[12] Audacity,[13] and Praat.[14] Because the data displayed are the magnitudes of the *overtones of the inverse of window size*, the graphical output depends dramatically on the window size. Figure 1 shows an example displayed using Wavesurfer[12]: The first word in ARCTIC database, speaker bdl, sentence a0008. The average pitch frequency is 125 Hz. The average pitch period is about 8 msec. By using a wide window, eg, 32 msec, the main features of the graph, the nearly horizontal features, are the pitch frequency and its overtones, see Figure 1(A). The intensity of each overtone is modulated by the timbre spectrum, with four formants at about 0.8 kHz, 2.0 kHz, 2.7 kHz, and 3.5 kHz. The timbre spectrum is deeply modulated by the fundamental frequency. By using a narrow window, eg 4 msec, the dominating feature is the power decay pattern within each pitch period, see Figure 1(B). The definitiveness of formant frequencies is substantially deteriorated. In both cases, the graphical display is a mixture of timbre information and pitch information.

Wavesurfer[12] also provides an option to display a curve of the timbre spectrum at any point of time. Figure 2 shows an example. The curve is again a mixture of timbre information and pitch information. It is very difficult to decipher formant frequencies from that curve.

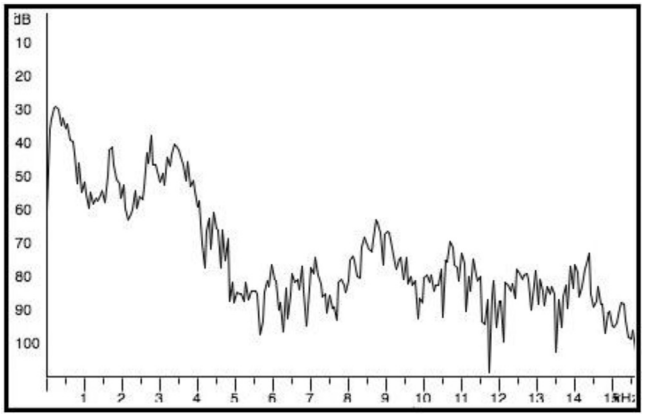

FIGURE 2. A spectrum section plot provided by Wavesurfer.

ELECTROGLOTTOGRAPH AND THE MECHANISM OF VOICE PRODUCTION

In 1957, French physiologist Philippe Fabre invented the electroglottograph[15] (EGG). The glottal closing instants can be precisely determined from the EGG signals.[16–18] The temporal correlation between the EGG signals and the voice signals is now well understood for all voice scientists: The voice signal in each pitch period starts as an intensive pulsation at the glottal closing instant. The voice signal is strong in the closed phase of the glottis, and weak in the open phase of the glottis. In the glottal closed phase, the voice signal decays exponentially. In the glottal open phase, the voice signal drops precipitously, often with some random noise added. See Figure 3.

The observed correlation between EGG signals and voice signals motivated Robert T. Sataloff to propose an analogy of human voice production with *hand clapping*.[19] According to Sataloff, "Sound is actually produced by the closing of the vocal folds, in a manner similar to the sound generated by hand clapping. . . . (T)he more frequently they open and close, the higher the pitch." This analogy is in line with the transient theory of human voice production.

The observed correlation between EGG signals and voice signals also motivated Ronald Baken to propose an analogy of human voice production with the *water-hammer effect*.[17] According to Baken, "The sharp cutoff of flow is particularly crucial, because it is this relatively sudden stoppage of the air flow that is truly the raw material of voice." Again, according to the water-hammer analogy, the sound wave starts at a glottal closing instant. The more frequently the glottis closes, the higher the pitch. This analogy is also in line with the transient theory of human voice production.

In the following, we make a careful quantitative analysis of what happens at a glottal closing, as shown in Figure 4.

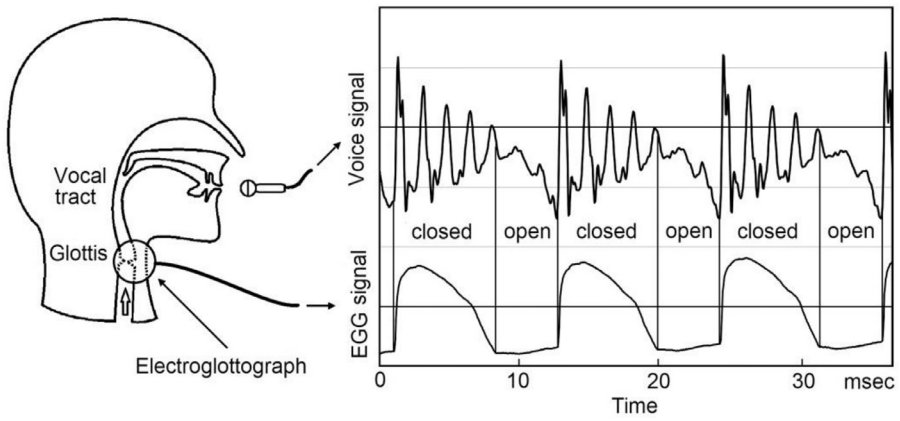

FIGURE 3. Temporal correlation of voice signals and EGG signals. Time delay due to sound wave propagation from the glottis to the microphone is corrected. The starting point of an elementary wave coincides with a glottal closing instant. The glottal opening instant has little effect on the voice. The speech signal in the closed phase is much stronger than that in the open phase.[18]

For simplicity, the vocal tract is represented by a tube of uniform cross section, roughly as the vocal-tract geometry of the vowel shwa [ə]. The length of the vocal tract, from the glottis to the lips, is about 160 mm. Density of surrounding air is represented by medium gray. Diluted air is represented by light gray. Compressed air is represented by dark gray. The small double arrows represent the particle velocity of air flow, with a typical value of 1 m/sec. The heavy arrow represents the speed of sound, with a typical value of 320 m/sec.

In step 0, in the open phase, air flows from the trachea to the vocal tract through the glottis. The density of air in the trachea and the vocal tract is the same as the surrounding air. In step 1, a glottal closing blocks the air flow from the trachea to the vocal tract. However, because of inertia (ie momentum), the air in the vocal tract keeps moving forward. The moving air creates a partial vacuum near the glottis. By partial vacuum, we mean a parcel of air which is diluted and with no particle velocity. There is a border between the moving air and the partial vacuum. According to the law of acoustics, the border is a wave front, which propagates with the speed of sound c, typically 320 m/sec, shown by the heavy arrow in step 2. As shown in step 3, within a time 160 mm/320 (m/s) = 0.5 millisecond, the wave front reaches the lips. Because the air in the vocal tract is diluted, the environment air rushes in to fill the partial vacuum with a reversed particle velocity, having a magnitude similar to the particle velocity of original airflow. As shown in step 4, the border between the moving air and the diluted still air is also a wave front, which propagates with the speed of sound c toward the glottis, shown by the heavy arrow.

The complete formant cycle is shown in Figure 5. The length of the vocal tract is denoted as L. In step 5, the wave front propagates toward the glottis with the speed of sound c. In a time

$$\tau = \frac{L}{c} \cong 0.5\,\text{msec}, \tag{1}$$

the wave front reaches the glottis, as shown in step 6. This time, the momentum of the moving air is pressing against the glottis. As a result, an air parcel near the glottis is compressed, as shown in step 7. The border between the moving air and the nonmoving compressed air is again a wave front, which propagates with the speed of sound c toward the open air, as shown by the heavy arrow in step 8. After yet another time τ shown in Eq. (1), the wave front reaches open air, as shown in step 9.

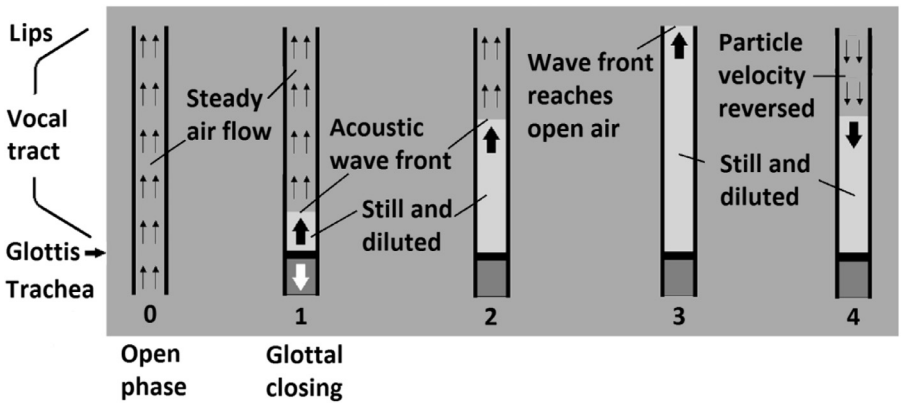

FIGURE 4. The initial steps of the closed phase in a pitch period.

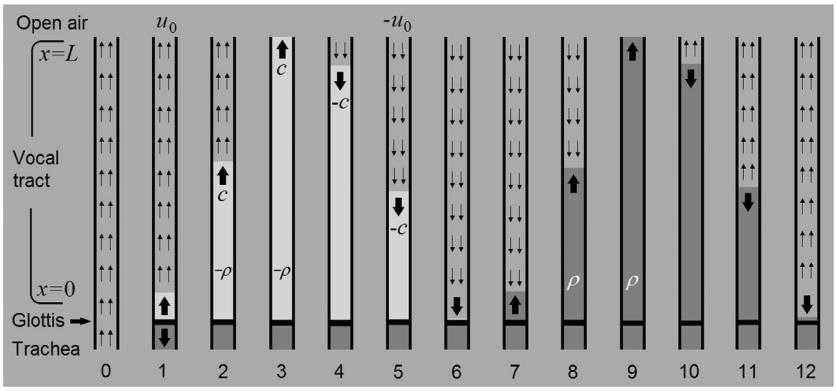

FIGURE 5. The complete process of a formant cycle in 12 steps.

Because the air in the vocal tract is compressed, it starts to move out into the open air, as shown in step 10. A border between the moving air and the compressed air takes place. As a wave front, it propagates with the speed of sound c toward the glottis, as shown by the heavy arrow in step 11. After yet another time τ shown in Eq. (1), the wave front reaches the glottis, as shown in step 12.

The configuration in step 12 is similar to the moment of glottal closing, step 1. The entire cycle repeats itself. However, because of radiation and friction, the particle velocity u_0 is reduced by a percentage. The total time of the cycle, including four phases, as shown in Eq. (2), is

$$T = \frac{4L}{c} \cong 2\,\text{msec}. \tag{2}$$

It represents a frequency

$$f_0 = \frac{c}{4L} \cong 0.5\,\text{kHz}. \tag{3}$$

It is the first formant of the vowel shwa [ə]. Because the waveform is square rather than sinusoidal, it has a series of overtones. The lowest overtones are $3f_0 = 1.5$ kHz and $5f_0 = 2.5$ kHz, approximately the second and third formants of the vowel shwa [ə]. The reduction of u_0 in each cycle means an exponential decay of intensity. Such an exponential decay is observed experimentally, see Figure 3.

The above explanation is not entirely new. It is the essence of the transient theory of voice production of Euler,[6] which was later quoted by Willis[7] and Lord Rayleigh[5]:

> If a single pulsation be excited at the bottom of a tube closed at one end, it will travel to the mouth of this tube with the velocity of sound. Here an echo of the pulsation will be formed which will run back again, be reflected from the bottom of the tube, and again present itself at the mouth where a new echo will be produced, and so on in succession till the motion is destroyed by friction and imperfect reflection.

Based on the above argument of a four-phase cycle, Euler obtained an expression of the resonance frequency,[6] Eq. (3). Because of the simple geometry in the vocal-tract model in Figures 3 and 4, an exact mathematical solution of the wave equation is derived, which accurately describes the transient process.[20] If the vocal tract has variable cross-sections, the formants are different, thus other vowels are produced.[20] Therefore, the correlation of voice signal and the EGG signal provides a more accurate understanding of the human voice production process, in line with the transient theory.

PITCH PERIOD VERSUS FUNDAMENTAL FREQUENCY

The correlation of voice signal and EGG signal provides another key for a basic concept in voice science. In the steady-state theory of voice, fundamental frequency is a starting point. For music instruments, the fundamental frequency can be defined scientifically.[9] However, in human voice, pitch varies constantly. In speech, the variation of pitch carries intonation. In tone languages, the variation of pitch in each syllable is a phonetic feature. Singing is a sequence of pitch alternation. And even for a single note in a melody, vibrato requires an uninterrupted variation of pitch. In both speech and singing, there is no scientifically valid definition of fundamental frequency from a physics point of view.

Based on EGG data, *pitch period* is a scientifically well-defined quantity from a physics point of view. Because of the Bernoulli force, the closing of glottis is extremely fast. Typically, the frequency of source signal of EGG is 200 kHz. The closing instant can be determined to accuracy better than 0.01 msec. The typical pitch period is of the order of 5–20 msec. In voiced sections, the time between two adjacent glottal closing instants can be physically defined to an accuracy of 0.1%. It is scientifically valid to define a *pitch period* as the time between two adjacent glottal closing instants.

If the pitch period does not change for an extended time, the inverse of the pitch period can be defined as the fundamental frequency. However, for human voice, it is seldom a physical reality.

Because for human voice, there is no valid definition of fundamental frequency from physics point of view, resonance as the profile of amplitudes of overtones of

fundamental frequency cannot be defined scientifically. On the other hand, the color of a vowel is carried by each transient wave triggered by a glottal closing. A *timbre spectrum* can be scientifically defined as the Fourier transform of each transient wave, as a feature of each pitch period.

METHOD OF PITCH-SYNCHRONOUS ANALYSIS

From the above discussions, it is understandable that pitch-synchronous analysis of voice signals could reveal more accurate pitch periods and timbre spectra. In fact, in the speech technology community, the advantages of pitch-synchronous processing of speech signals have been known for several decades. However, in spite of numerous efforts,[21–23] no reliable methods to segment the voice signals into pitch periods and subsequent processing have been found.

Here we show that the simultaneously acquired EGG signals provide a scientific foundation of pitch-synchronous segmentation, see Figure 6. Typically, in the derivative of EGG signals, the closing instant exhibits a strong and sharp peak. Because of the time delay due to the distance from the glottis to the microphone, which is typically 20− 30 cm, the peak of the dEGG signal is 0.6−1.0 msec ahead of the strong peak in the voice signal. It is usually in the glottal open phase, where the voice signal varies weakly. It is the best place to implement an ends-matching procedure, see Figure 7.

In general, the pcm values of the two ends of a pitch period do not match. As shown in Figure 7(A), the starting point of a pitch period, ***S***, does not equal the end point ***E***. In order to perform a Fourier analysis, the values of the waveform at the two ends should be made equal. To achieve this, a small section of voice signal in the previous period, ***P*** = ***CS***, is provided. At the end of the pitch period, another small section of the voice signal, ***Q*** = ***DE***, is also provided. The number of points in those transitional sections is equal. For example, with a sample rate of 32 kHz, for a signal of pitch frequency 160 Hz, the number of signal points in a pitch period is 200. Typically, the number of points K in the transitional section is about 10% of the pitch period. Here we should have K = 20. A new section ***R*** is then generated

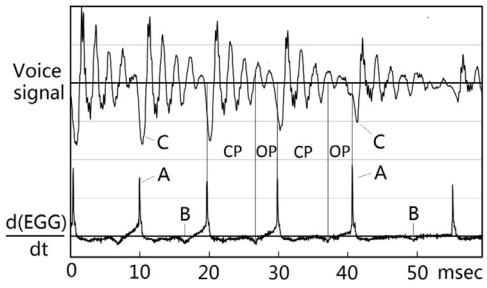

FIGURE 6. Using the sharp peaks in dEGG as pitch marks. Here, A is the peak of glottal closing, B is the peak of glottal opening. CP is the closed phase, OP is the open phase.

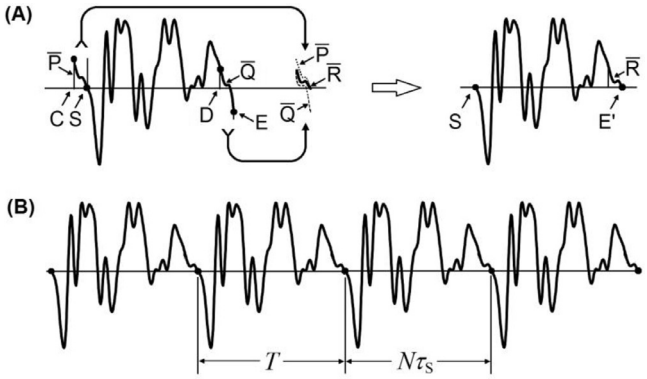

FIGURE 7. The ends-matching procedure.

as a graded linear combination of the two original signals, ***P*** and ***Q*** as:

$$R(k) = \frac{K-k}{K}P(k) + \frac{k}{K}Q(k). \tag{4}$$

Then, in the signals of a pitch period, the original values of Q(k) are replaced by the new values of R(k). From the above equation, it is clear that near the beginning of the transitional section, k=0, the value of R(k) equals the original value Q(k). Near the end of the transitional section, k=K, the value of R (k) equals the value of P(k), that is, the value at the starting point S. Therefore, after this minor modification, the value at the end of the pitch period, ***E'***, equals that at the starting point ***S***. In between, the new values represent a gradual transition from P(k) to Q(k). Because the transitional section is in the glottal open phase, the variation is small, the replacement of a transitional section should cause little disturbance to the timbre spectrum, see Figure 7(B).

In general, the number of pcm points in a pitch period, $s(n)$, does not equal to a power of 2, FFT is difficult to apply. The original Fourier analysis is applied by multiplying the voice signal with sine and cosine functions,

$$\begin{aligned} a(m) &= \frac{2}{N}\sum_{n=0}^{n<N} s(n)\cos\frac{2\pi nm}{N}, \\ b(m) &= \frac{2}{N}\sum_{n=0}^{n<N} s(n)\sin\frac{2\pi nm}{N}. \end{aligned} \tag{5}$$

On modern computers, those floating-point computations are sufficiently fast. The amplitude spectrum is then calculated as a combination of the sine and cosine components, as shown in Eq. (6),

$$c(m) = \sqrt{a(m)^2 + b(m)^2}. \tag{6}$$

The number of independent Fourier amplitudes is one half of the number of points in each pitch period. In the current example, $N = 200$. Therefore, the range of m is 100. According to the sampling theorem, the maximum frequency is 16 kHz. The number of frequency points is 100. Therefore, each frequency point is an integer multiple of 0.16 kHz, or 160 Hz.

To make a display of the spectrum on a graph, the values of Fourier components have to be interpolated to the number of pixels, which is a fixed number determined by the display. On the other hand, the number of available Fourier coefficients depends on the pitch, or pitch period. Take an example, if the vertical size of the display for a 16 kHz spectrum is 320 points, the 100 spectral points must interpolate into 320 points.

A standard method for such interpolation is to use the Whittaker-Shannon formula. However, sometimes it generates too much ringing. To improve it, the Whittaker-Shannon formula is truncated with a Gaussian envelop, as shown in Eq. (7),

$$f(x) = e^{-\kappa x^2} \frac{\sin \pi x}{\pi x}. \tag{7}$$

with κ is 0.16 and the value of x is limited from −4 to +4. See Section 5.7.2 in Reference (20).

The value of the pitch frequency strongly influences the accuracy of the timbre spectrum, as expected from the uncertainly principle.[20] The frequency resolution limit is related to the inverse of the pitch period. Because of the interpolation process, if a formant frequency lies between two overtones, then interpolation process can generate a curve, where the peak is located at the formant frequency. However, if there are two formants between two overtones, there is no way to detect both. Therefore, to improve the frequency resolution of the timbre spectrum, a lower pitch is preferred. Female speakers typically have a higher pitch. However, at the end of phrases, many female speakers tend to lower the pitch. Therefore, high resolution timbre spectrum can also be achieved for female speakers.

For unvoiced sections, the voice signal is segmented with an interval equal to the average pitch period in voiced sections, then matches the ends using Eq. (4) and performs Fourier analysis using Eq. (5).

Pitch-synchronous segmentation can also be performed from voice signals. This is especially important for voice signals without simultaneously acquired EGG signals or the EGG signals are faulty. For details, see Section 5.4 of Reference 20. In the evaluation package, such a program is provided.

GRAPHICAL DISPLAY OF PITCH PERIODS AND TIMBRE SPECTRA

Based on the algorithm described above, a software package with an intuitive user interface Psyncs (short for Pitch Synchronous Spectrogram) is developed. For academic users, a free evaluation version can be obtained from Columbia Technology Ventures.[24]

The entire package is included in a directory, temporarily named PSS. By copying this directory to a Windows computer and open it, the content of the directory appears, as shown in Figure 8(A). The doc subdirectory contains the user's manual and the text of the recordings. The exe directory contains four executables and some image files. The tmp directory contains temporary files generated during the execution. The icon of the application Psyncs implies that the software is dedicated to human voice and is based on its production mechanism.

For testing and demonstration, the package includes two sets of standard speech files, derived from the Carnegie-Melon University's ARCTIC databases. Published in 2002, it is now the most widely used test samples for speech technology. Included in the package are the first 20 sentences spoken by a male speaker bdl, and a female speaker slt. Both are in US English. Originally it has simultaneously acquired EGG signals. Here, only the voice signals are included. Pitch marks are derived from the voice signals. Phoneme label display is an essential feature of the pitch-synchronous spectrogram software. Although the original ARCTIC databases contain phoneme labels, the accuracy is not good. The 40 phoneme label files included in the software package were manually inspected and corrected.

By double-clicking the icon Psyncs, an input panel appears, se shown in Figure 8(B). The first field is Speaker. There are two speakers in the demonstration sets, bdl, and slt. Any user can add voice signal files of new speakers. The second field is sentence identification. Currently, the

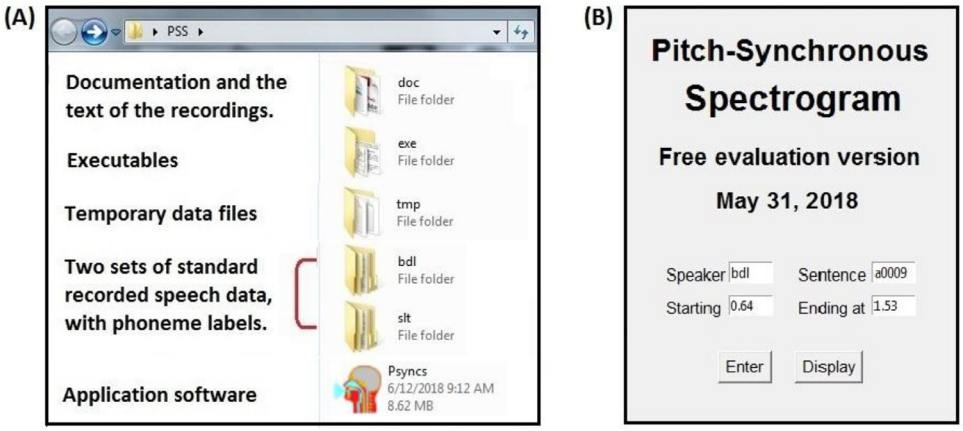

FIGURE 8. A. the content of the root directory of the software package. B. the first screen after clicking the application icon Psyncs, showing four input fields.

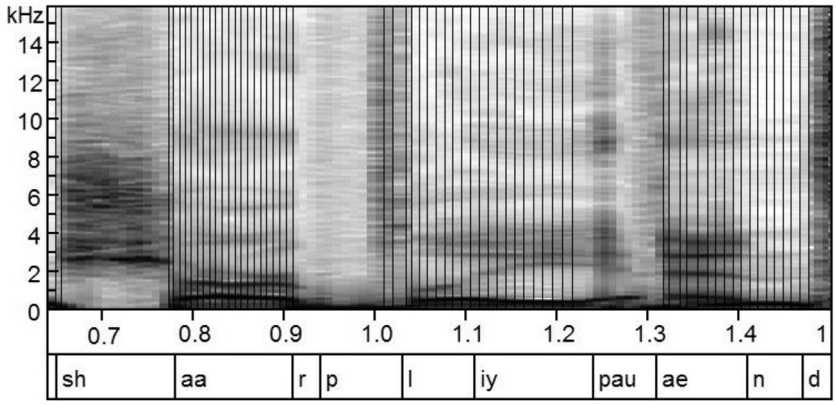

FIGURE 9. The pitch-synchronous spectrogram. In voiced sections, each vertical line represents a glottal closing instant. The timbre spectrum is displayed as gray levels. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

sentence identification for the ARCTIC databases is a0001 to a0020. The user can define their own sentence names. The third and the fourth fields are starting time and ending time, in second(s). A nice view can be generated by displaying speech signals with duration of about 0.25−1.5 seconds.

After clicking the "Enter" and the "Display" buttons, a spectrogram appears, as shown in Figure 9. Here are two words ".. sharply and .." in sentence a0009 by speaker bdl. Sections with voiced phonemes are characterized by an array of vertical lines, each represents a glottal closing instant. The time between two adjacent vertical lines is the pitch period. For each pitch period, the timbre spectrum in kHz is displayed as gray levels in that column. As shown, pitch information and timbre information are cleanly separated. There are no artificially imposed window size and window functions. The unvoiced sections do not have glottal closings, thus there are no vertical lines, but are segmented with a constant interval.

To display the timbre spectrum of a pitch period, place the cursor at that period, then left click. Figure 10 shows an example, displayed in a linear scale of spectral power. As shown, because pitch information is eliminated, the spectrum is clean and sharp. A lot of details are disclosed.

Another feature of human voice is the variation of instantaneous power inside each pitch period. In traditional spectrogram, by using a very narrow window size, for example around 1 msec, the decay of instantaneous power, from the highest value near the glottal closing instant to the middle of the glottal open phase is apparent, as shown in Figure 11. The pattern of decay of instantaneous power is a pitch period is an important feature of human voice, such as the role of closed quotient or open quotient.[18] By making a right-click, the power evolution pattern in this pitch period is displayed, see Figure 11. As shown, in the glottis closed phase, the power decays slowly and exponentially. After the glottis opens, power decay dramatically accelerates.

Figure 12 shows the timbre spectra of four vowels from the pitch-synchronous spectrogram. As shown, the graphics of the spectra exhibits rich details with accuracy and cleanness.

Figure 13 shows the timbre spectra of four consonants. For unvoiced consonants, the frequency range is 16 kHz. As shown, for fricative [sh], the spectral power concentrates in the range of 3 kHz and 7 kHz. For fricative [f], spectral power concentrates between 12 and 15 kHz.

Figure 14 shows four power decay patterns of different vowels. The spectral power of vowel [ae] is mostly above 1 kHz, its decay in the closed phase is fast. In some cases,

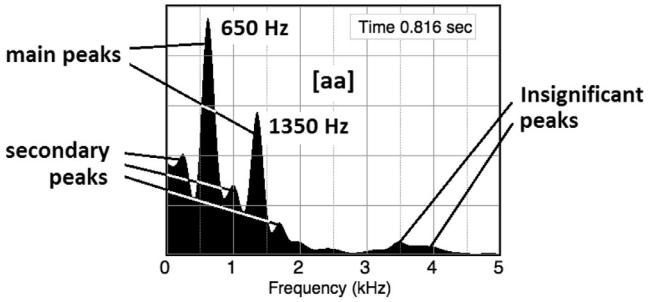

FIGURE 10. The timbre spectrum of a pitch period.

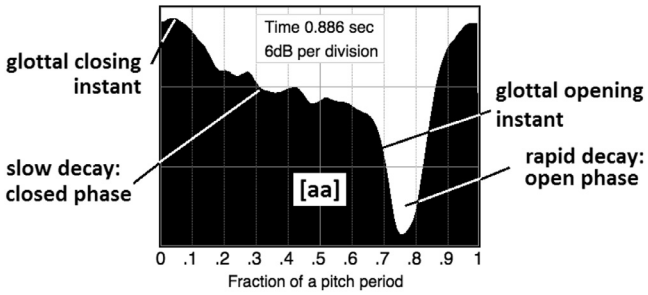

FIGURE 11. Power evolution pattern in a pitch period.

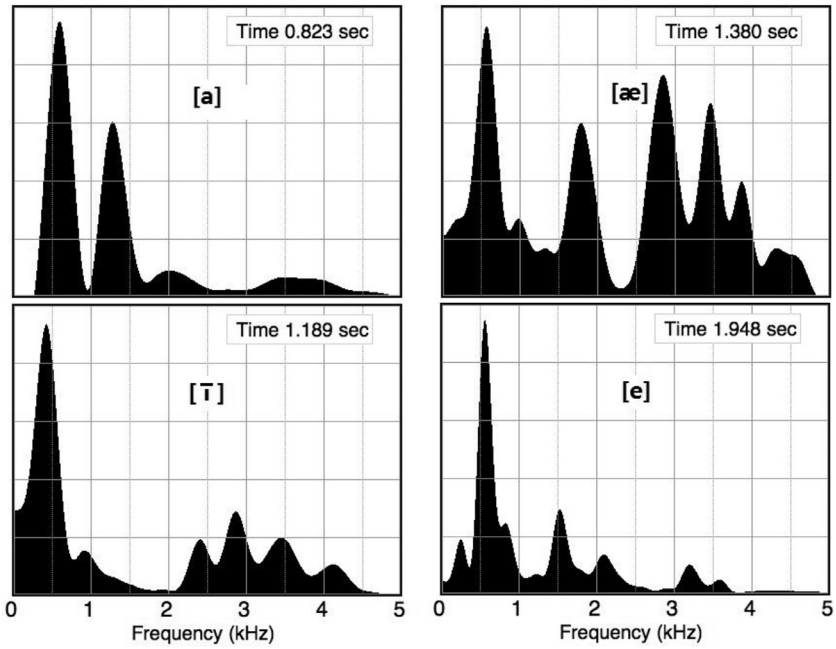

FIGURE 12. Timbre spectra of four vowels, by speaker bdl in ARCTIC databases.

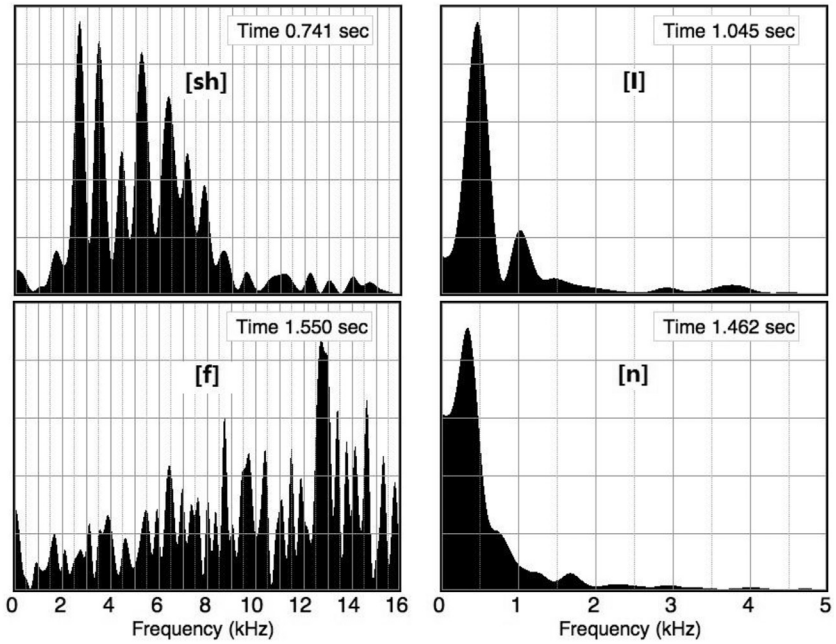

FIGURE 13. Examples of timbre spectra of several consonants.

an excitation at the glottal opening moment is observed. A rich body of information is to be studied and analyzed.

LIMITATIONS

A limitation of the pitch-synchronous spectrogram, comparing to the traditional spectrogram, is that it can only handle human voice uttered by a single person. For general sound, instrumental music, and choral music, it does not work. For general sound and music signals, the traditional pitch-asynchronous spectrogram is still the preferred tool.

CONCLUSIONS

Simultaneously acquired voice and EGG signals indicate that each glottal closing triggers a decaying elementary wave in the vocal tract. A superposition of those elementary waves constitutes voice. Based on that concept and using EGG data, a pitch-synchronous voice signal processing

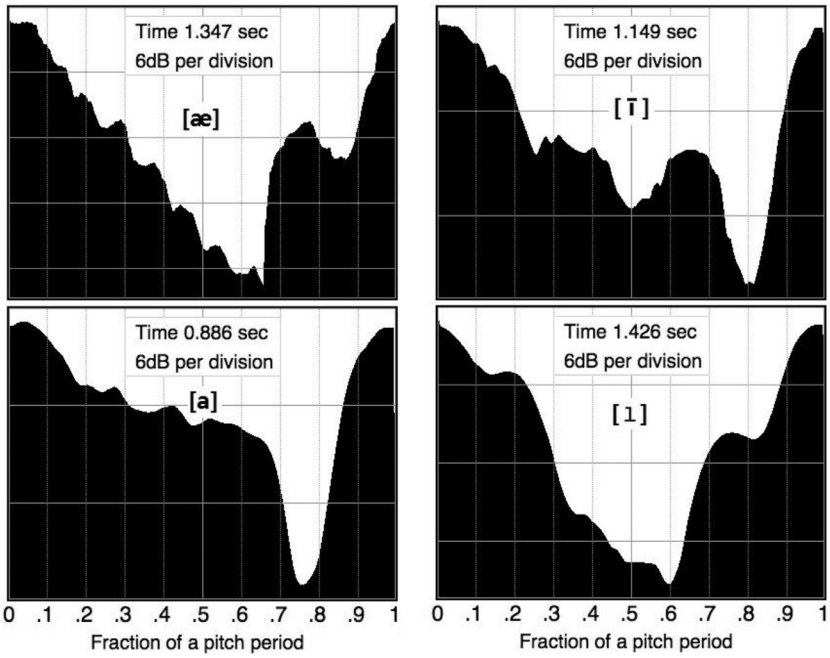

FIGURE 14. Examples of power decay patterns of several vowels.

method is developed. The voice signal is first segmented into pitch periods, then the two ends are equalized. Fourier analysis is applied to obtain the timbre spectrum of each pitch period. High resolution visual display of timbre spectrum is generated. The power evolution pattern in each pitch period is also displayed.

Acknowledgments

The authors wish to thank Richard Lissemore for inspiring discussions.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.jvoice.2019.01.009.

REFERENCES

1. Fulop S. *Speech Spectrum Analysis*. Berlin Heidelberg: Springer Verlag; 2011.
2. Chiba T, Kajiyama M. *The Vowel, Its Nature and Structure*. Phonetic Society of Japan; 1942. (Japanese edition), 1958 (English edition).
3. Scripture EW. *The Elements of Experimental Phonetics*. New York: E. Scribner's Sons; 1902. The entire book is digitized by Google and available online.
4. Ladefoged P. *Elements of Acoustic Phonetics*. Second Edition University of Chicago Press; 1995.
5. Rayleigh JWS. *The Theory of Sound*. London: McMillan; 1894. The entire book is digitized by MSN and available online.
6. Euler L. *Dissertation physica de sono*. Euler Archive; 1727:E002.. translated and annotated by Ian Bruce.
7. Willis R. On the vowel sounds, and on reed organ pipes. *Trans Camb Phil Soc*. 1830;III:231.
8. Wheatstone C. Reed organ-pipes, speaking machines, etc. *London and Westminster Review*. 1837;28:27.
9. Helmholtz HLF. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York: Dover Publications; 1954. Original German edition published in 1863. Translated by H. Margenau into English.
10. Fant G. *Acoustic Theory of Speech Production*. Mouton De Gruyer; 1970.
11. Rabiner LR, Schafer RW. *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall; 1978.
12. See the homepage for Wavesurfer, www.speech.kth.se/wavesurfer.
13. See the homepage for Audacity, www.audacity.sourceforge.net.
14. See the homepage for Praat, www.fon.uva.nl/praat.
15. 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*. 1957;141:66–69.
16. Baken RJ. Electroglottography. *J Voice*. 1992;6:98–110.
17. Baken RJ. A review of laryngeal function for voice production. In: Sataloff Robert, ed. Third Edition*Professional Voice*. 1, Plural Publishing; 2005:237.
18. Miller DG. *Resonance in Singing*. Inside View Press; 2008.
19. Sataloff R. *The Human Voice*. Scientific American; 1992:108, 1992.
20. Chen CJ. *Elements of Human Voice*. World Scientific Publishing; 2016.
21. Miller JE, Mathews MV, David Jr EE. Pitch synchronous analysis of voiced sounds. *J. Acoust Soc Am.*. 1961;33:179.
22. Hess WJ. A pitch-synchronous digital feature extraction system for phonemic recognition of speech. *IEEE Trans ASSP*. 1976;24:14.
23. Medan Y, Yair E. Pitch synchronous spectral analysis scheme for voiced speech. *IEEE Trans ASSP*. 1989;37:1321.
24. Open webpage innovation.columbia.edu/technologies/CU18357, then follow the instructions.