Research Article

Journal of Electrical Electronics Engineering

Multi-Dimensional Spectral Process for Cepstral Feature Engineering & Formant Coding

Ahmad Z. Hasanain[1*], Muntaser M. Syed[2], Veton Z. Këpuska[3], Marius C. Silaghi[4]

[1]*College of Engineering at Al-Lith, Umm Al-Qura University, Kingdom of Saudi Arabia*
[1,2,3,4]*Electrical & Computer Engineering Department, Florida Institute of Technology, United States of America*

[*]Corresponding author

Ahmad Zuahir S Hasanain, Umm Al-Qura University, Al Lith, KSA 28434
azhasanain@uqu.edu.sa

Submitted:02 Aug 2022; **Accepted:**09 Aug 2022; **Published:**15 Aug 2022

Citation: *Hasanain, A., Syed, M., Kepuska, V., Silaghi, M. (2022). Multi-Dimensional Spectral Process for Cepstral Feature Engineering & Formant Coding. J Electrical Electron Eng, 1(1), 01-20.*

Abstract

The fundamental frequency feature is essential for Automatic Speech Recognition because its patterns convey a paralanguage and its tuning normalizes other speech features. Human speech is multidimensional because it is minimally represented by three variables: the intonation (or pitch), the formants (or timbre), and the speech resolution (or depth). These variables represent the hidden states of the local glottal variation, the vocal tract response, and the frequency scale, respectively. Computing them one by one is not as efficient as computing them together, so this article introduces a new speech feature extraction approach.

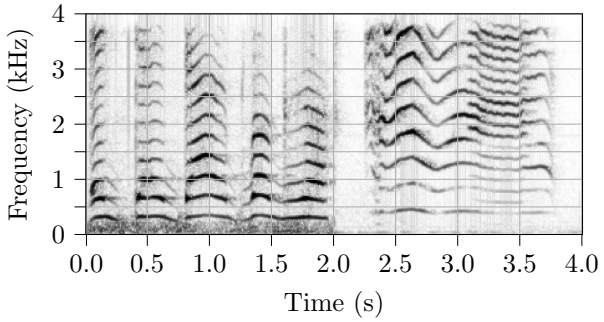
The article is introductory; it focuses on the basic concepts of our new approach and does not elaborate on all applications. It demonstrates that the unit of a cepstral value, which is a spectral value of spectrums, is a unit of acceleration since its discrete variable, the quefrency, can be expressed in Hertz-per-microsecond. The article shows how to produce refined voice analysis from robust estimates and how to reconstruct speech signals from feature spaces. And it concludes that the pitch track of the new approach is as good as two open-source pitch extractors.

Combining multiple processes, attenuating background noises, and enabling distant-speech recognition, we introduce the Speech Quefrency Transform (SQT) approach as well as multiple quefrency scales. SQT is a set of frequency transforms whose spectral leakages are controlled per a frequency-modulation model. SQT captures the stationarity of time series onto a hyperspace that resembles the cepstrogram when it is reduced for pitch track extraction.

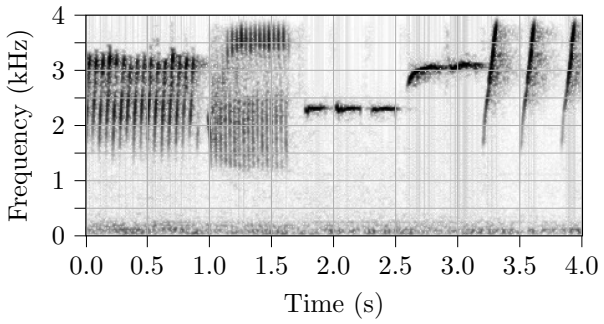
Key Words: Feature Engineering, Pitch Track, Cepstral Coefficients, Speech Synthesis

1. Introduction

At first glance at Figure 1, one can notice the parallel curves in the spectrogram of the human voice (Figure 1a) but not in the bird chirp (Figure 1b). These salient curly harmonics of the voice render a hidden state that appears contentious when connecting the dots. The spectrogram graphs show snaps of spectrums, which is the energy distribution of the frequency components, versus time. The contrasts of the graphs in this article were adjusted for printing such that the higher the energies, the darker the pixels, but the energy scales were omitted, and the color scheme can be reversed when printed on monitors. The human speech has two features that are visible in the spectrograms: the pitch and the harmonic intensities, whose patterns can represent abstract concepts and boost intelligence. In order to have artificial agents processing or understanding spoken languages naturally, it is crucial to realize a mathematical representation for speech that is



(a) Human Babble [1]



(b) Canary Chirp

Figure 1: Spectrograms of Multi- and Mono-Resonance Communication Systems

accordingly attuned since human intelligence and language may appear tangled up during development.

Voice producers flap in response to changes in internal air pressure, air molecules are compressed and released periodically, and the pulse shape makes speech signals transmittable through air. The linearly-spaced curves in the spectrogram are the outcome of a periodic time signal that is locally stationary. Hence, the periodicity and the period shape fluctuate slowly when compared to the sampling rate. In the spectrograms, the more the signal is locally stationary, the sharper the curves are. The time distance between two adjacent compressions (i.e., bursts, pulses, or cycles) is the wave period (T_0), measured in seconds per cycle (1/Hz). The wave-interval is the reciprocal of the minimum frequency shift between two harmonic curves, as in Equation 1. This minimal shift is the speech fundamental frequency (f_0). Per context, it is also the pitch and the frequency carrier. However, being in an air medium as its communication channel, the signal's actual periodicity is measured in meters per cycle. The λ_0 and υ in the equation are the corresponding wavelength and the speed of sound in the channel. Although the variables are time variants, the υ is usually assumed to be constant, but the temperature, humidity, and wind speed, all of which slightly affect υ, are not constant along the air paths from the speech producer (vocal folds, cords, or glottis) to the speech receivers (eardrums' cochleas, microphones, and acoustic beamformers).

$$T_0 = \lambda_0/\upsilon = 1/f_0 \text{ (seconds per cycle)} \quad (1)$$

In Figure 1b, the transitioning of the birdsong f_0 is constrained in the producer's hyper-coordinates, but the f_0 observations are projected onto the two-dimensional spectrogram. The projection onto the periodicity space is non-linear since the f_0 teleports in the spectrogram as though two frequencies (such as 1 kHz and 4 kHz) are identical because there are unaccounted independent axes. For example, the fundamental waveform of the canary is visually rotated around a time-variant axis parallel to the time axis, and its perimeter path renders a visual effect of cylinders that are visible. Assuming the bird's mono f_0 was traveling with a constant angular velocity in a polar coordinate, the inferred radius of a pictured cylinder is about 1 kHz and centered at 2 kHz.

Similarly, the infant voice in Figure 1a appears with a deeper voice during an emotional outburst, between Second 3.25 and 3.6. The event is noticeable in the figure and also in the audio playback. The tone-change phenomenon, which usually happens during puberty, doubles the fundamental interval and folds up the spectral code bandwidth. This creates the speech resolutions. Deep and high human voices are not represented equally in a telephone bandwidth (within 4 kHz). Additionally, there have been several frequency scales, and the variable scaling of the spectral bandwidth appears to have been one of the main challenges in Automatic Speech Recognition (ASR). In order to normalize the speech features, the speaker's pitch must be considered during the feature extraction process.

Per the juxtaposition of the two spectrograms, the human voice has a fundamental waveform, whose shape transformed in the figure at a relatively slow pace, and each of whose parallel spectral curve produced a component. The spectral energies of the harmonic components are mainly the speech features. The harmonic components can also be called timbre or overtone series. They function as the frequency-modulating signal and the vocal tract response, denoted B in this work for modeling the speech system. According to Stefanatos et al. [2], the human perception of speech is similar frequency demodulation. If the parallel curves were the openings of window blinds, a few shaded patterns would appear behind the blinds. The shaded patterns are the formants, and their mixtures' variations compose the phonemes. A phoneme is a distinctive sound, resulting from convolving the multitone signal with the formant system, as illustrated in Figure 2. The human speech consists of these components, which are conveying the hidden shape of the spatial cavities of the vocal tract (the nasal and oral cavities). The collective shape of the tract is a system through which the molecules' excitations of the fundamental waveform pass. The output of the

modulating system is the speech signal, which, due to its local stationarity, consists of recognizable time units.

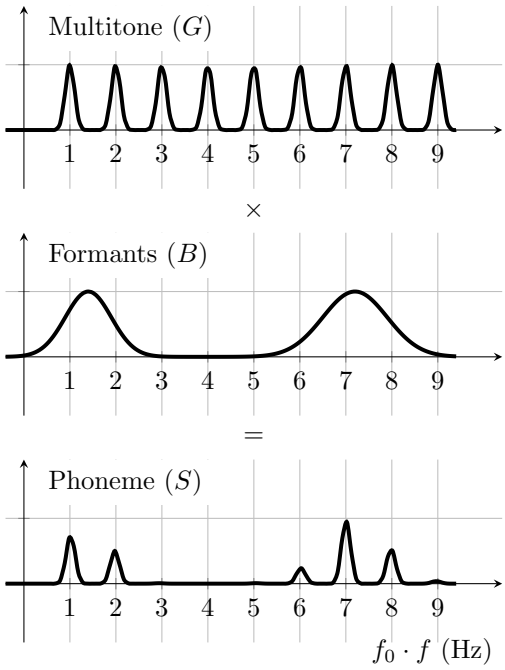


Figure 2: Phoneme Representation in the Spectrum

Unlike the birdsong but like many mammals' sounds, human speech consists primarily of the multitone signal that travels through a multi-resonance vocal system, and it is characterized by local periodicity as sketched in Figure 3. Equivalently, in the frequency domain, as shown in Figure 2, the glottal train of the multitone signal is multiplied by a Gaussian Mixture Model of the formants. The periodicity, hence the fundamental frequency f_0, controls the frequency spacing between the elements of the speech code in the channel bandwidth, and high speech components are attenuated. Two commonplace approaches to pitch and speech feature extractions are modulation-based and are applied to the frequency domain. The Fast Fourier Transform, albeit invaluable for several applications, can complicate some speech processing tasks such as f_0 tracking. Also applying the Fourier and/or the Cosine Transforms twice for the quefrency domain does not normalize the speech features. Bogert [3] coined the term quefrency, which was derived from the term frequency; compare qu-e-fr-ency and fr-e-qu-ency. The quefrencies are sometimes regarded as the inverse of frequencies. The next section (Section 2) goes through a brief review of the related literature. However, for the SQT methodology in Section 3, the quefrencies are simply frequencies of frequencies. Then, an SQT algorithm for pitch extraction is defined in Section 4, tested in Section 5, and analyzed in Section 6. The analysis section (Section 6) discusses further findings. The article is concluded by a summary, implications, and future work.

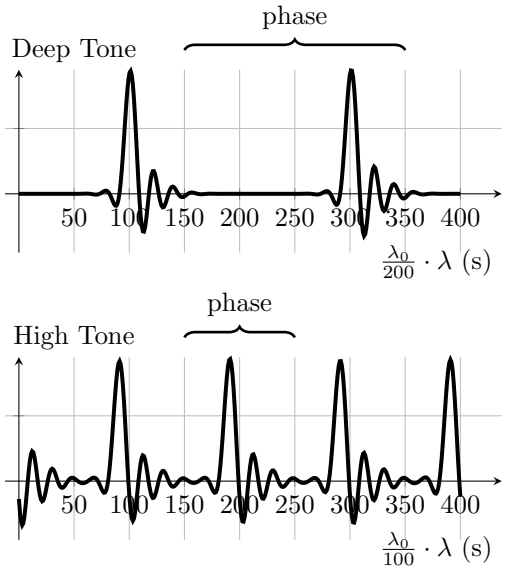


Figure 3: Pulse Trains

2. Related Work

The pitch track represents the glottal state, and it can be extracted from speech signals in several ways. The points of strengths of some approaches complement the points of weaknesses of other approaches, and so the features of their methods can be combined in some applications. In this overview, a few examples are highlighted for their unique specifications, but there are many methods to extract the pitch track in literature. Even though several techniques had existed for pitch extraction according to Rabiner et al. [4], it was still one of the most computationally demanding modules according to Hess [5]. The leading methods include Normalized Correlation Function (NCF), Pitch Estimation Filter (PEF), Cepstrum Pitch Determination (CPD), and Mel-Frequency Cepstral Coefficients (MFCC). Generally, the pitch track methods operate in the temporal, spectral, and cepstral domains. "It is about transforming data from passive to active, from static to dynamic - transforming data into insight. Now, all of this demands a new approach to information technology from the approaches of the 80s or the 90s," Carly Fiorina said at Oracle OpenWorld 2004.

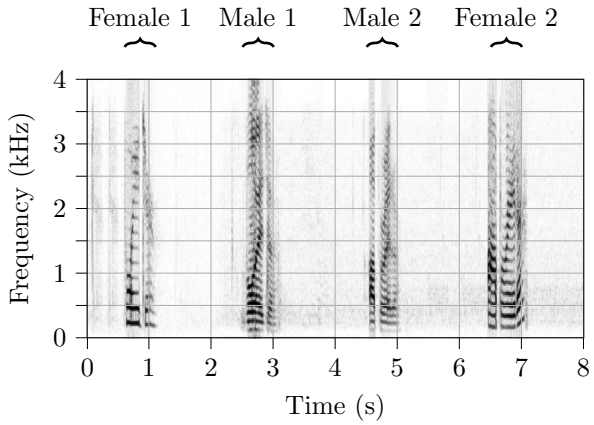
In the time domain, the signal can be matched with its lagged versions using auto-correlation, which is one of the basic methods for pitch extraction, as shown by Rabiner et al. [4]. Since the speech signal is assumed to be stationary, it can be compared with itself, and the self similarity is measured high when the lag matches the wavelength. Another general way to estimate the pitch is from the number of the zero crossings or the sign flipping. In the frequency domain, the spectral components of the signal are obtained first, as has been shown in the spectrogram. Since the speech signals have overtones, the frequency components can indicate the fundamental frequency. A common spectral method is the harmonic

product spectrum, which De La Cuadra et al. [6] showed effective in adjusting acoustic instruments, but it is very prone to noise, and so it may not be used for speech pitch tracks. Moreover, the instantaneous frequency and phase can be combined with the time analysis to extract the pitch track, as shown by Kawahara [7], Charpentier [8], although reacquiring high frequency resolutions. Pitch Estimation Filter (PEF) uses the frequency domain, and its performance is in Section 5. In the cepstral domain, which is commonly used as the power spectrum of the logarithm of the power spectrum according to Bogert et al. [9], Noll [10], Oppenheim and Schafer [11], the pitch track and its overtones become separated from the fromant features, and cepstral filtering, also known as liftering, is applied to identify the pitch track from the overtone lookalikes. Examples of this approach are Cepstrum Pitch Determination (CPD) and Mel-Frequency Cepstral Coefficients (MFCC). The equation for calculating the cepstrum is defined as $\mathcal{F}^{-1}log|\mathcal{F}\{\bullet\}|$ where $\mathcal{F}\{\bullet\}$ denotes a forward Fourier transformation, $e^{-j\theta} = cos(\theta) - j \cdot sin(\theta)$, and $j = \sqrt{-1}$ according to Lathi and Green [12]; hence, Equation 2.

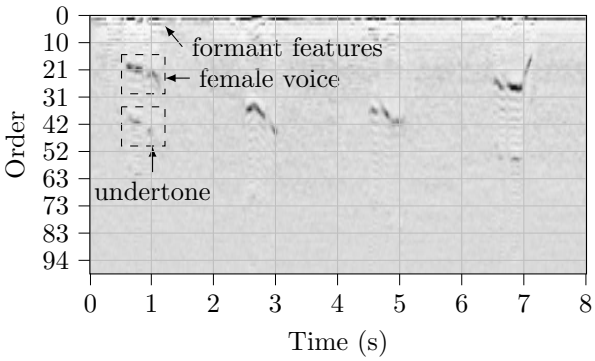
$$p[n] = \frac{1}{c}\sum_{m=0}^{c-1} e^{j2\pi\frac{nm}{c}}\ log\Big|\sum_{u=0}^{c-1} e^{-j2\pi\frac{mu}{c}} s[u]\Big| \qquad (2)$$

The main concern of the pitch track methods is differentiating between the pitch and its first overtone. The tones that are one unit octave apart from the true pitch track are lookalikes since they have in-common components. Using the Normalized Cross-Correlation (NCF) method, Talkin [14] showed three tones that two of which were lookalike noises. Also the methods by Talkin [14], Ewender et al. [15], Atlas and Janssen [16], Li and Atlas [17] were not instantaneously able to filter the pitch from its partial harmonic components. Similarly, the overtone ambiguity is present in the Mel-Frequency Cepstral Coefficients (MFCC), which is considered one of the best speech feature extractions as it has been proven robust in various speech applications according to Ganchev et al. [18]. To visualize the pitch ambiguity in the MFCC features, consider the MFCC cepstrograms of four utterances in Figure 4, which depicts the first hundred MFCC features versus time. A cepstrogram is a spectrogram of spectrograms, and quefrency is the independent variable of the cepstrum just as frequency is so of the spectrum. In the first cepstrogram (Figure 4b, the utterances were isolated and did not have background noise. However, in the second cepstrogram (Figure 4c, the utterances were coupled by child cries. In the first graph, the overtone noise appeared in the two words that were uttered by female voices. For example, there were two tones at 0.5 second. In the second MFCC graph, the background noise appeared distorting the pitch track.

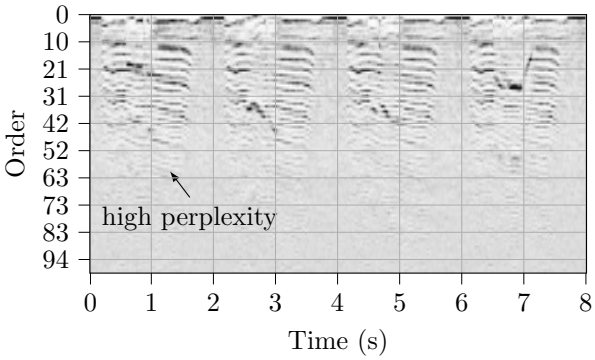
There are two strategies to mitigate the drawbacks: combining and smoothing. In the first strategy, the pitch ambiguity are circumvented by putting together the pitch tracks of various methods according to Moorer [19]. Since the various features of the methods may be partially independent, putting them together may reduce the collective number of their blind spots. For example, Zahorian and Hu [20] combined observation candidates from more than one pitch detection approach in the "Yet Another Algorithm for Pitch Tracking" (YAAPT). The other workaround is smoothing. Since the



(a) Spectrogram of Utterances, from Kepuska [13]



(b) MFCC Features of Utterances

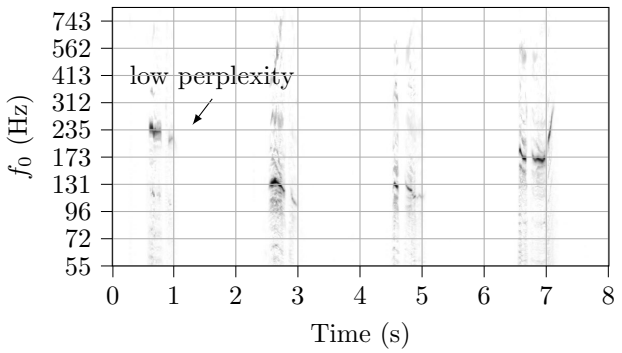


(c) MFCC Features of Utterances and Infant Noise

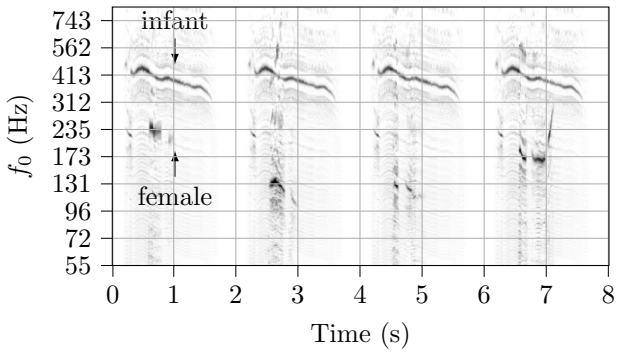
Figure 4: Cepstral Perplexity of the MFCC Pitch Tracks (Generated in Python by Librosa)

pitch track is likely not to have abrupt changes due to the stationarity, multiple observations may stabilize the reading and filter out the outliers. Although the temporal averaging has been used to address the main concern by Kawahara et al. [21], it obviously lowers the detection quality, especially at the edges of the voice activity intervals. The original characteristics of the data get corrupted when mean filters are applied on the pitch track. Generally, the bulk of the algorithms in the literature perplexes with the overtones because of the f_0 harmonic characteristic. Sometimes, human perception regards them more similar than other tones. However, the ambiguity in f_0 should not exist in the acoustic frontend since human beings can easily differentiate between the speech depths.

"It is about transforming data from passive to active, from static to dynamic - transforming data into insight," Carly Fiorina said at Oracle OpenWorld 2004. The MFCC cepstrograms in Figure 4 are not as good as the SQT cepstrograms in Figure 5 for several reasons. First, SQT has a flexible quefrency scale while the MFCC series is fixed. Consequently, SQT can provide higher cepstral resolutions than MFCC does. For example, the MFCC resolution in Figure 4b is lower than the SQT cepstral resolution in Figure 5a. Second, SQT may satisfactorily lifter out the overtone noise where MFCC struggles with the overtone noise of the high voices. Therefore, SQT



(a) Utterances



(b) Utterances With Infant Noise

Figure 5: SQT Cepstral Features (SQT, Python)

appears to be more suitable in crosstalk applications than MFCC does. For example, the separation between the simultaneous pitch tracks in Figure 4c is not as good as the feature separability of SQT in Figure 5b. The SQT methodology attitudes the resolution of the speech noise. It is analogous to the human ability to tune to one speaker and have the ambient voice disregard (or suppressed and blurred). Third, SQT is not prone to white noise as MFCC is, and so it is more suitable than the cepstral approaches for distance pitch extraction. When the speaker is a few meters away from a low quality microphone, the signal-to-noise ratio can become small even when the background noise is not speech. Because the SQT approach is designed based on a human speech model, its cepstrograms are good for pitch track extraction in the case of distance speech recognition when compared with the cepstral approaches. For these reasons, the SQT cepstrograms are novel and distinctive. The SQT flexibility in terms of quefrency scales and cepstral resolution makes it practical for a wide range of applications, as it enables custom quefrency distribution, multi-speaker, and distant (far field) pitch track extractions. The SQT approach is noise resilient because it is based on sound theorems and numerical proofs as shown in the next section.

3. Approach

The acoustic perception is receptive to frequency-modulated signals. From Lathi and Green [12], the amplitude of a single tone can be modulated into a Direct Current (DC) value when it is multiplied by a similar tone, as in the frequency demodulation derivation in Equation 3. The filter is similar to the tone when their frequencies are matched ($f_1 = f_2$) and their phase is synchronized ($\phi_1 = \phi_2$). The DC value ($2B$) is obtained by taking the arithmetic average of the result of that multiplication. As shown in the MFCC cepstrogram (Figure 4), the quefrency domain is a sophisticated speech feature space since it separates between the voice feature and the tract feature, but there is a need for a new model to overcome the cepstral drawbacks. Liftering the cepstrum to locate the pitch track after its construction is hardly useful. The cepstral filtering has to be integrated before hand in the spectrograms, whose frequency bands can be wide and narrow. In our new approach, the bands of the spectrograms respond to the desired cepstral sinusoide filters, which measure cycle accelerations. In a multi-dimensional SQT method, responsive spectrograms are generated by Speech Quefrency Transform and are then reduced in dimensionality to produce the SQT cepstrogram.

Quefrency is simply a measurement of acceleration. This can be derived directly from the relevant definitions when they are applied on a spectral impulse train. In simple words, a frequency shift changes the cycle velocity. Some people may argue that the cepstrum looks like the time space as it is commonly used for lossy compression, and

$$
\begin{aligned}
\text{Filter} \cdot \text{Tone} &= 2A\ cos(\alpha f_1 + \phi_1) \cdot 2B\ cos(\alpha f_2 + \phi_2) \\
&= (A\ e^{-i\alpha f_1} e^{-i\phi_1} + A\ e^{i\alpha f_1} e^{i\phi_1})(B\ e^{-i\alpha f_2} e^{-i\phi_2} + B\ e^{i\alpha f_2} e^{i\phi_2}) \\
&= \Big|_{f_1 = m f_2} \quad AB\ e^{-i\alpha f_1 (1-m)} e^{-i(\phi_1 - \phi_2)} + AB\ e^{i\alpha f_1 (1-m)} e^{i(\phi_1 - \phi_2)} \\
&\qquad + AB\ e^{-i\alpha f_1 (1+m)} e^{-i(\phi_1 + \phi_2)} + AB\ e^{i\alpha f_1 (1+m)} e^{i(\phi_1 + \phi_2)} \\
&= 2AB\ cos(\alpha f_1 (m-1) + \phi_2 - \phi_1) + 2AB\ cos(\alpha f_1 (m+1) + \phi_2 + \phi_1) \\
&= \Big|_{\phi_1 = \phi_2,\ m=1,\ \&\ A=1} \quad \underbrace{2B}_{\text{DC Value}} + 2B\ cos(2\alpha f_1 + 2\phi_1)
\end{aligned} \tag{3}
$$

thus its unit must be seconds, but this is not necessarily true. Some spaces retain similar characteristics; however, they have different units. For example, the second derivative of $f(x) = x^3$ is monotonically increasing just like the original function, yet $f(x)$ and $f''(x)$ have different units. However, if the unit of $F(x)$ is meters per second and the unit of x is second, then the unit of $F(F(x))$ is meters per second squared, which is a unit of acceleration. While frequency measures the velocity (cycles per second), quefrency measures the acceleration of the cycles. Additionally, while the speech time samples have positive and negative values oscillating around a zero mean, the speech frequency samples are usually represented with positive energy magnitudes, each of which has an angular phase whose range is in $[0, 2\pi)$ radians (rad). However, the spectral components can have negative values. For example, the real part of the energy vector is considered negative in the third and fourth quarters: $[\pi, 2\pi)$ rad.

The spectral oscillating function of the SQT approach is attainable by applying the well known theories of Fourier, Stone-Weierstrass, and Nyquist on the frequency domain. Per Fourier, the quefrency filter osculates in the frequency domain; in other words, positive and negative frequency bands (or banks) have to be designated for each cepstral filter. Based on Stone-Weierstrass Theory, the time-bounded spectral signals are approximated by sums of polynomials. Therefore, for the real quefrency filter, the cosine function can be approximated by its first M exponential terms, as in Equation 4. Note that the exponential terms are mixture of Gaussian windows, as shown in Equation 5. Figure 6 demonstrates that there is negligible difference between the cosine function and its approximation when the independent variable is in the range $(0.5, M)$. Based on Nyquist Theory, the SQT approach requires that the width of the main lobe of the window function be less than or equal to $f_0/2$. This is because the spectral oscillation requires an alternating sign.

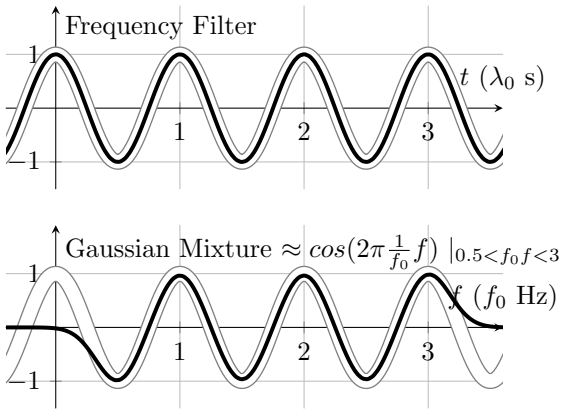


Figure 6: A Cosine Function & Its Third-Order-Gaussian Approximation

$$
\cos(2\pi\lambda_0\ f) \approx \sum_{m=1}^{M} e^{-(f - m\cdot f_0)^2/\sigma^2} - e^{-(f-(m-0.5)\cdot f_0)^2/\sigma^2} \tag{4}
$$

$$
w_g[u] = \begin{cases} \frac{f_0}{2f_s} \cdot e^{-\frac{1}{2}\cdot\left(u\frac{f_0}{f_s} - 1\right)^2}, & \text{if } u \in \{1, 2, \cdots, \lceil 2f_s/f_0 \rceil\} \\ 0, & \text{otherwise} \end{cases} \tag{5}
$$

The cosine function is directly applicable in the time domain but may not be directly applicable in the frequency domain. One possible way to construct a spectral sinusoide filter is designing an all-pass filter whose phase curve approximates a cosine function. Another hypothetical way is to design an all ripple frequency response. However, both of them are not straightforward as we found out. The practical method is to obtain one component at a time. The computational complexity of the multi-dimensional method increases linearly, while the other two methods factorially, which is faster than exponentially. The quefrency model, therefore, should be constructed by a set of temporal sinusoidal filters whose frequency responses assemble spectral sinusoidal filters when they are aggregated. In either methodology, the window function has to vary per quefrency and should be applied on the frequency filters to assemble the quefrency filters. The proof in the previous section is shown using Gaussian window because of its simplicity, but the window does not have to be Gaussian. As long as the conditions are met, several other windows can be applied. For

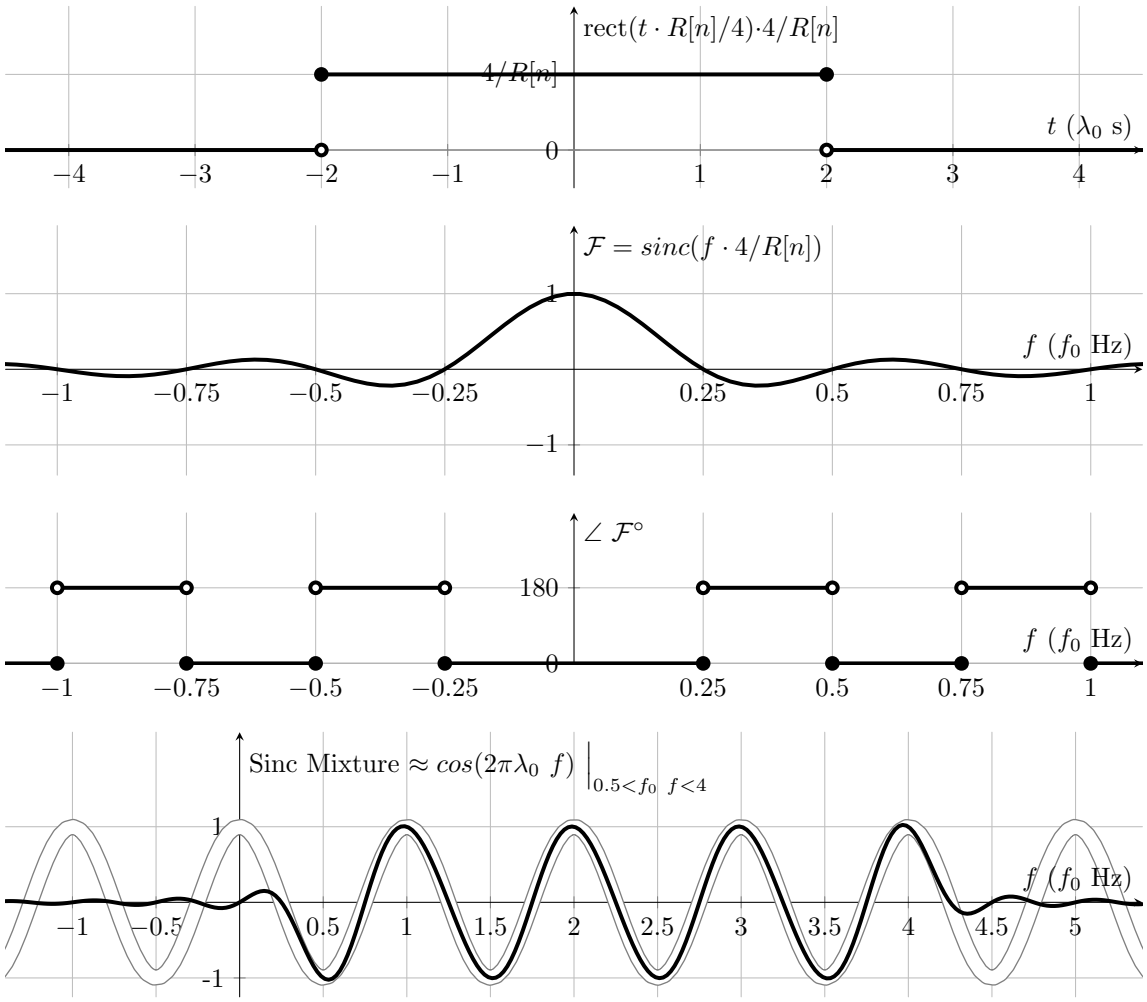


Figure 7: Sinc Frequency Banks

example, because Parseval's Theorem is defined on rectangular windows, as in Equation 6, extracting the SQT features using rectangular windows gives accurate energy extraction and signal reconstructions. However, the SQT matrix of rectangular windows is two times bigger than the SQT matrix of Gaussian windows. This is because the optimal main lobe width of the Gaussian window is $f_0/2$. Meanwhile, the optimal main lobe width of the rectangular window is $f_0/4$, as shown in Figure 7 and Equation 7. In the figure, the sinc mixture is depicted with $M = 4$, and its equation converges to a cosine function in the interval between 0.5 and M as $M \to \infty$. For that reason, the rectangular window of SQT is defined in Equation 8.

This section presented the SQT approach, which constructs quefrency filters by utilizing the spectral leakage of the windowing filters. Note that, once the signal is in frames (convolved by sliding), the speech sequence is re-sampled from the sampling rate to the frame rate (R_s), which is measured in frame-per-second. Because speech exists once it is in a physical medium, the original unit of its cycles is in meters (or miles), and so quefrencies are accelerations just as frequencies are velocities (cycles per second). The section covered the Gaussian and rectangular windows, but other windows may be applicable as well. For example, The Dolph-Chebyshev window may distribute the spectral noise uniformly according to Ykhlef et al. [22], so its pitch tracks may be

$$\mathrm{RMS}[t] = \sqrt{\frac{1}{c}\sum_{u} s^2[u,t]} \approx \sqrt{\sum_{m} B^2[m,t]} \tag{6}$$

$$cos(2\pi\lambda_0 f)\,\Big|_{0.5<f_0 f<M} \approx \sum_{\tau=1}^{2M}(-1)^{\tau}\cdot sinc\big((f-\tau/2)\cdot\lambda_w[n]\big) \tag{7}$$

$$\Big|\ sinc(x) = \frac{sin(\pi x)}{\pi x}$$

$$w_r[u] = \begin{cases} \frac{f_0}{4f_s}, & \text{if } u \in \{0,1,\cdots,\lceil 4f_s/f_0\rceil - 1\} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

more noise resilient than other windows. Nevertheless, to center the measurements, it is crucial that the number of cycles inside the window be even. While the rectangular window has speech recognition applications, the Gaussian window appears sufficient for extracting the pitch track. Therefore, this article continues with the Gaussian window. Figure 8 shows the filtering implications of the SQT approach when it encounters undertone and white noises. These two types of noises are likely to be canceled out in their mean filter values because they tend to have equal powers (or energies) at the positive and negative sub-bands of the quefrency filters. To show how to utilize the SQT approach using the multi-dimensional method, the next section defines three quefrency scales and an algorithm that includes further cepstral filtering.

4. Procedures

Speech signals are continuous with an infinite sampling rate (f_s), but they become time-discrete when sampled or digitized. The sampling rate (f_s) is 8 kHz in the telephonic narrowband applications and is 16 kHz in the multimedia wideband. The sampled speech series is then again discretized to time frames, as shown in Figure 9. In the figure, the frame rate (R_s) was lowered only for simplification, but there has to be some overlapping between the adjacent frames. From Frame 1 of the Onward utterance, the pitch is roughly 200 Hz since the wavelength (period or cycle interval) appears at 5 ms. The sampling rate of the auditory and the visual perceptions

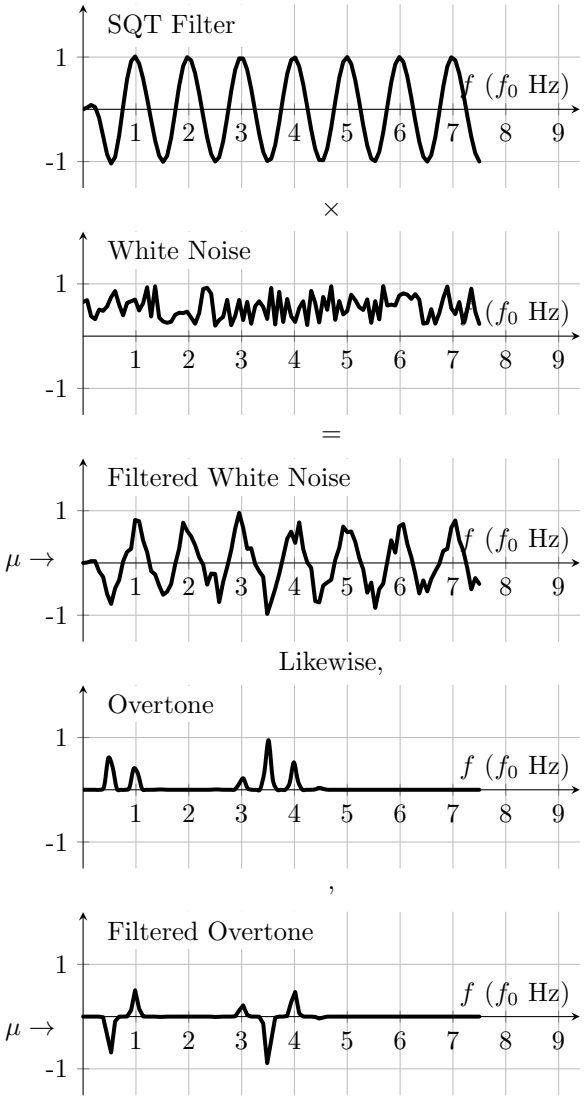
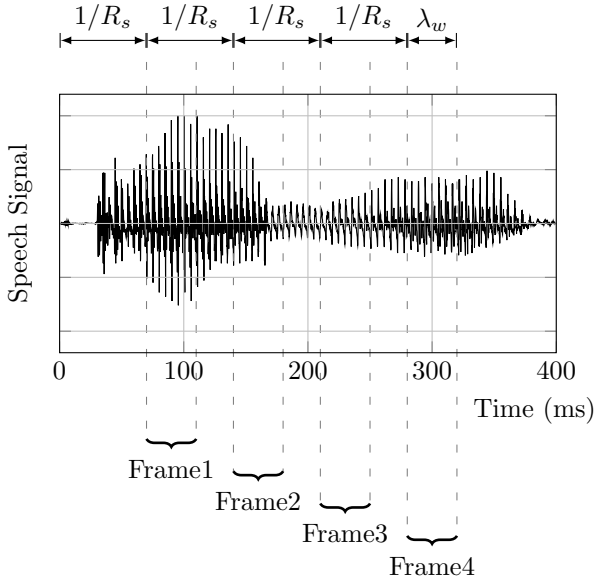
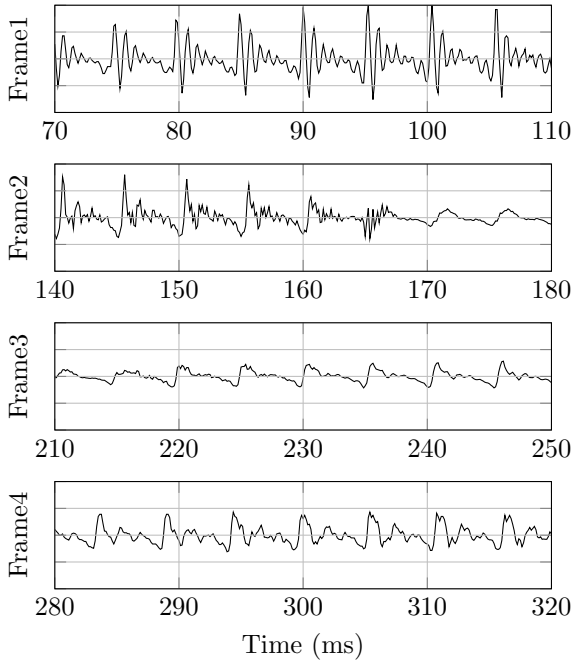


Figure 8: SQT Filtering

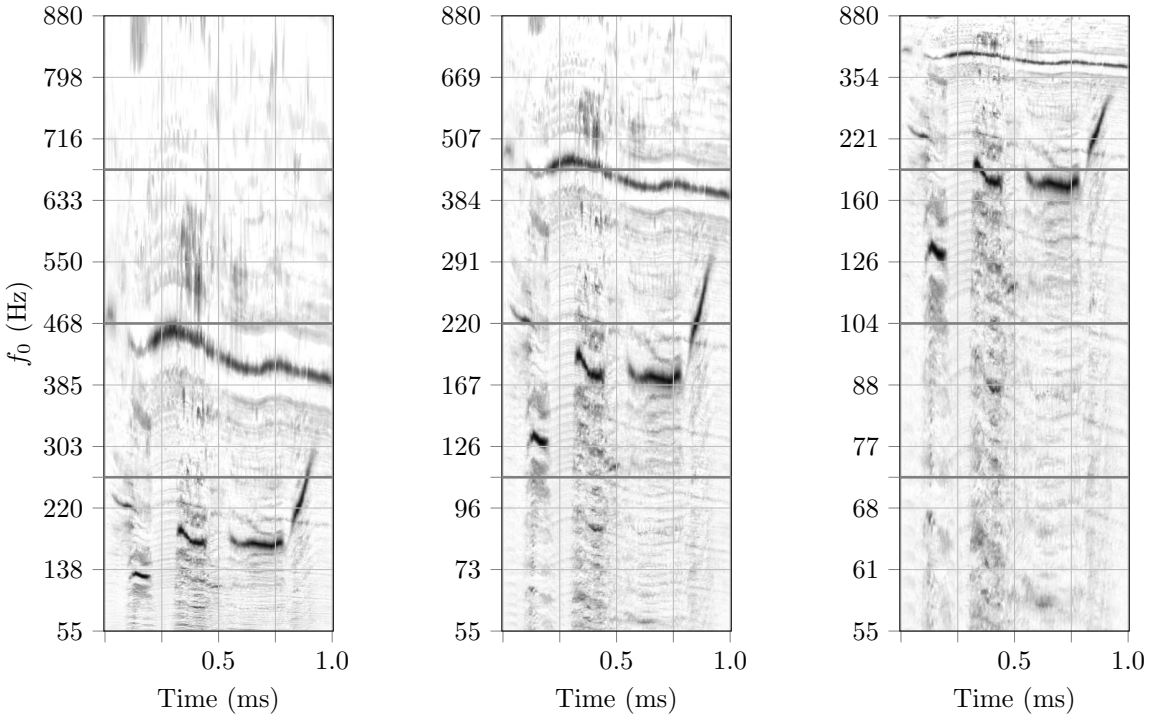


(a) Pulse Code Modulation (PCM) Series



(b) Speech Frame Matrix

Figure 9: Window Sliding by Frame Rate R_s



(a) Linear (b) Geometric (c) Reciprocal

Figure 10: SQT Scales

$$
\begin{aligned}
R_{\text{Lin}}[n] &= \{r : r = f_{min} + (f_{max} - f_{min}) \cdot n/N, n \in \{0, 1, \cdots, N\}\} \\
R_{\text{Geo}}[n] &= \{r : r = f_{min} \cdot (f_{max}/f_{min})^{R_{\text{Lin}}[n]/N} \cdot n/N, n \in \{0, 1, \cdots, N\}\} \\
R_{\text{Rec}}[n] &= \{r : r = 1/(1/f_{min} + (1/f_{max} - 1/f_{min}) \cdot n/N), n \in \{0, 1, \cdots, N\}\}
\end{aligned}
\tag{9}
$$

may be relevant. For example, 60 frames per second (fps) is considered a high frame rate standard for video, and it may be so as well for the audio frame rate (R_s). Although f_s is proportional to the frequency range, it does not have a direct relationship with its in-frame resolution. The frequency resolution relies on the window interval (span or length). In other words, one needs to increase the window length to measure lower frequencies. Human hearing may perceive frequencies down to 20 Hz according to psychoacoustic experiments by Wölfel and McDonough [23]. The spectral range of the audio instruments is above 50 Hz according to Rabiner and Gold [24], and the speech volume at 50 Hz is considered to be the lowest level in the standardized acoustic loudness contours according to Standard of International Organization for Standardization [25].

Three quefrency scales are applied in Figure 10 and defined in Equation 9; they are: Linear-Space, (R_{Lin}), Geometric (R_{Geo}), and Reciprocal (R_{Rec}). From the bandwidth of the three pitch tracks in the figure, one can see why the geometric scale is better than the linear and reciprocal spaces. It is because the SQT geometric scale distributes the cepstral pixels moderately. On the other hand, the resolution of the linear space is biased toward high quefrencies while the resolution of the reciprocal space is biased toward low frequencies. It is commonly believed that human perception differentiates between deep voices better than it differentiates between high voices. This is similar to both the reciprocal scale and the MFCC scale. It is worth to note that the cepstral resolution (N) in the SQT figure (Figure 10c) is ten times higher than the cepstral resolution in the MFCC figure (Figure 4b). However, the information gain of the pitch track can be increased when the quefrency scale represents the speakers' distribution regardless of the listeners' perception. For that reason, we prefer the geometric scale to the linear and the reciprocal scales ($R[n] = R_{\text{Geo}}[n]$).

When the spectral resolution is M pixels and the cepstral resolution is N pixels, the complex SQT transformer of the multi-dimensional method is defined in Equation 10, and its real part is depicted in Figure 11, where the $\lceil \bullet \rceil$ operator rounds up the $\bullet$ value to the nearest integer. In the SQT

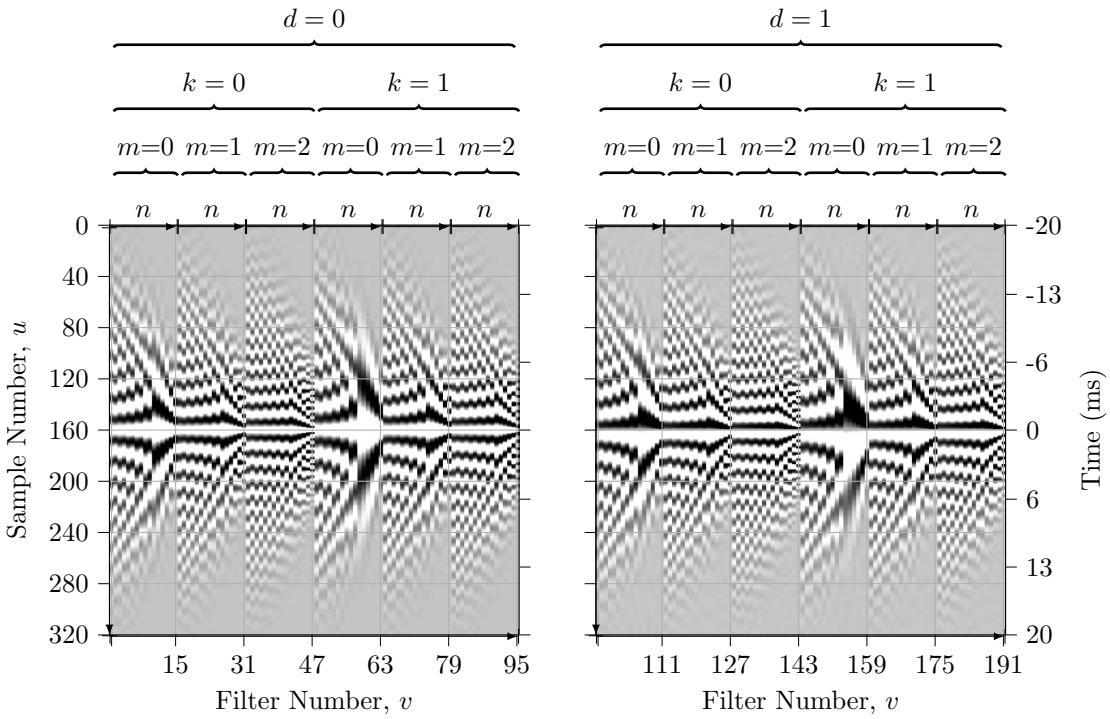
$$T_{4D}[u,v] = \begin{cases} \frac{e^{-\frac{1}{2}\cdot\left(u\frac{R[n]}{f_s}-1\right)^2}}{2f_s/R[n]} cos(f_i(u-\frac{c}{2})+k), & \text{if } f_i \le \pi \ \&\ |u-\frac{c}{2}| \le \frac{f_s}{R[n]} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$k = (\lfloor v/N/M/2 \rfloor \bmod 2)\ \pi/2, \quad f_i = 2\pi(m - d/2 + 1)\ R[n]/f_s$$

$$d = \lfloor v/N/M \rfloor \bmod 2, \quad m = \lfloor v/N \rfloor \bmod M, \quad n = v \bmod N,$$

$$v \in \{0, 1, \cdots, 4MN-1\}, \quad u \in \{0, 1, \cdots, c-1\},$$

$$\text{and } c = \lceil 2f_s/R[0] \rceil$$



(a) Even Part (b) Odd Part

Figure 11: SQT Matrix

matrix (T), all harmonic filters of an f_0 value are measured in the same window. Additionally, the maximum window length is defined by the minimal quefrency ($R[0]$) in the equation, as has been explained: The interval of the sliding frame corresponds to the lower limit of the fundamental frequency since f_{min} is reciprocal to the number of cycles in the window. At least two f_0 cycles must be enclosed in the frame to be able to detect the local stationarity of speech signals, as in Frames 2-4 in Figure 9. From the demodulation perspective, the quefrency of filter has to be at least double the frequency of the modulated signal. That is, based on Nyquist theorem, at least two cycle observations are necessary for a reliable detection. The minimal window interval required to detect f_0 without aliasing is $\frac{f_0}{2}$.

Feature engineering steps of an SQT processare summarized in the block-diagram of Figure 12. These steps obtain the pitch track and an $M \times N$ responsive spectrogram, with both of which, the speech series may be approximated (or reconstructed) back. SQT is applied by matrix multiplication with the Frame Matrix, shown in Figure 9: Frames $\times$ T. In the multi-dimensional SQT space, two binary variables (d and k) expand the speech dimensions into the real ($d = 0$) and imaginary ($d = 1$) parts and the positive ($k = 0$) and negative ($k = 1$) parts. In the diagram, the two dimensions are reduced once they are extracted. The Distance values are obtained from the real and imaginary parts using Pythagorean theorem ($|B| = \sqrt{B_{\Re}^2 + B_{\Im}^2}$). The Rectify values are obtained by

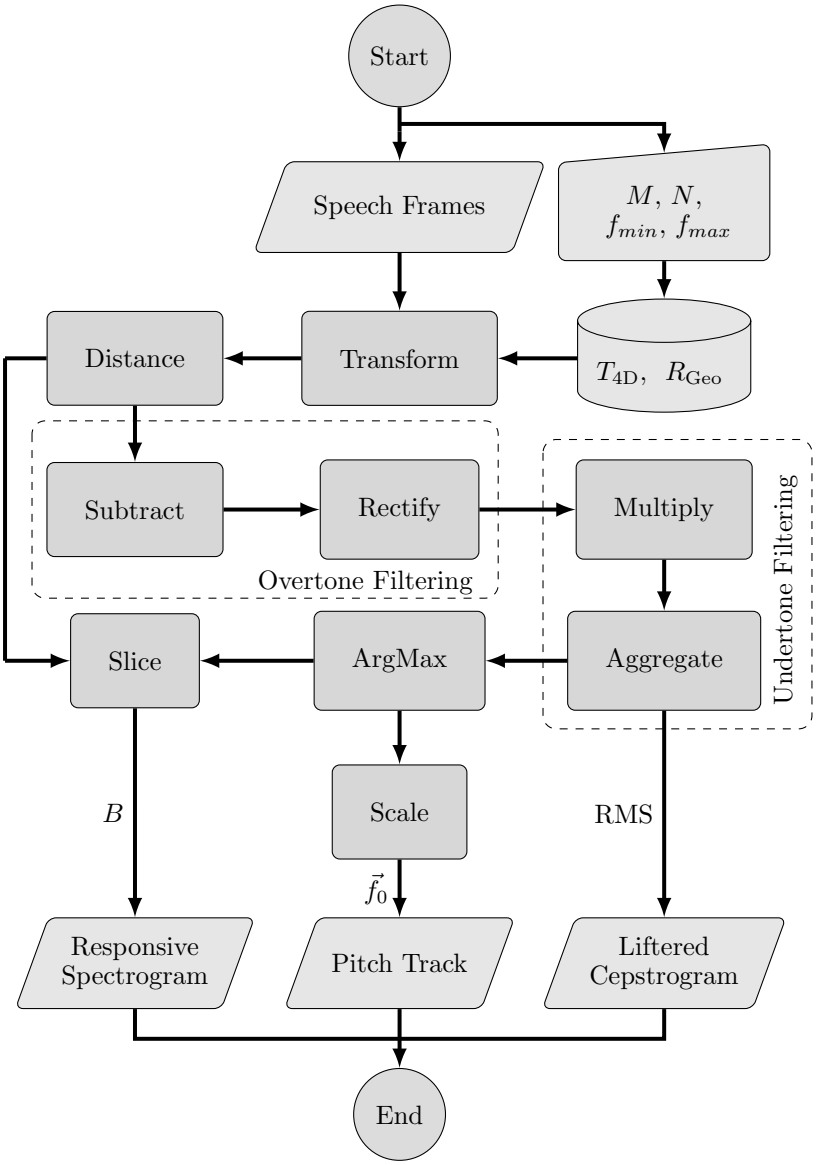
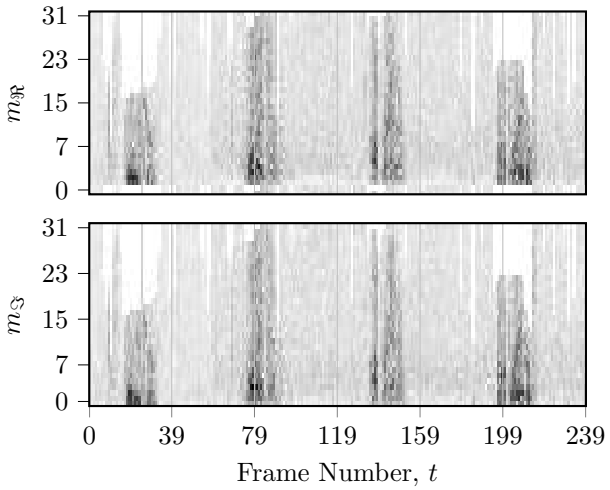


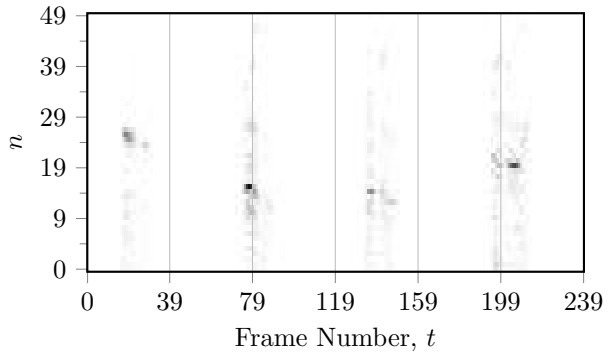
Figure 12: Hyperspace Reduction Paradigm

subtracting the negative part from the positive part and zeroing the negative values: $\max(B_+ - B_-, 0)$. It can optimally be preceded with a linear mean filter along the time axis to smooth the high harmonic components. The remaining dimensions are the spectral (m) and cepstral (n) dimensions. The Multiply operator is a non-linear multiplication filter that is applied along the m-axis. It multiplies the adjacent harmonic terms to filter the overtone noise. In this process, it is applied on the responsive spectral domain to lifter the cepstrogram just as the window is usually applied on the temporal domain to filter the spectrogram. Finally, to obtain the pitch track, The m and n dimensions are reduced by Aggregate, which returns the sum along the m dimension, and ArgMax, which returns the n index along the last dimension. The n index, which is defined in the quefrency scale, is then used to slice the Distance matrix. Additionally, to obtain the filtered cepstrogram, the Aggregate operator is applied along the n-axis. More operators may be added to this minimal example for the applications other than the pitch track extraction.

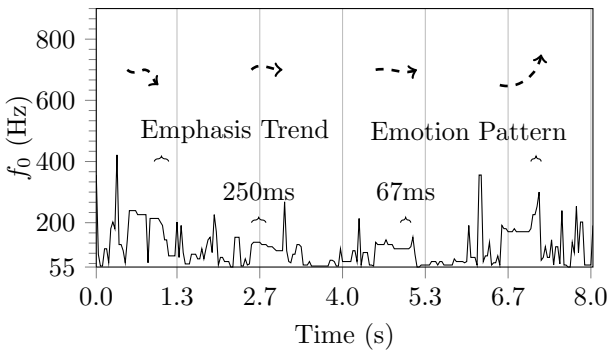
Figure 13 shows the outputs of an extraction example whose N is 50, M is 32, and $\hat{R}_s$ is 30. The first graph (Figure 13a) depicts the real and imaginary parts of its complex responsive spectrogram (B). The cepstral resolution shown in the cepstrogram of the second graph (Figure 13b) is much lower than the one in Figure 5a. The third graph (Figure 13c) plots the pitch track by applying



(a) Complex Responsive Spectrogram



(b) Liftered Cepstrogram



(c) Pitch Track

Figure 13: Outputs of a Complex Extraction Example

the ArgMax operator on the low resolution cepstrogram. The edges of the pitch track appears sharp, as premised. Another output example is in Figure 14, whose estimator is compared with the original frames of Figure 9. This time, the quality of the pitch extraction was tested by reconstructing the original speech signal using the complex spectra. A linear-interpolation formula for the reconstruction is simplified in Equation 11, which embodies a modular function (Equation 12) and a limiting compressor with a normalized threshold: $\alpha \in [0, 1)$. To use this formula, the actual frame rate (R_s) during the feature extraction must be updated by $R_s = f_s/\lfloor f_s/\hat{R}_s \rfloor$, when $\hat{R}_s$ is estimate by end-users. Speech utterances sound natural when they are carried on the f_0 pattern. For instance, the syllables of verbs are emphasized differently than those of nouns. Consequently, the pitch track is one of the most important speech feature in Natural Languages Processing. According to Albert Mehrabian [26], the nonverbal human tone carries shades of meanings. The emotions, such as happiness and sadness, may correlate with the pitch pattern according to Green et al. [27]. As annotated in Figure 13c, the pitch tracks can highlight some parts of speech, and the alerting acceleration may be crucial for Wake-Up-Word recognition.

5. Results

In the previous sections, we showed the output of the approach by high cepstral resolutions; however, the pitch track can be extracted by lower time costs. This section shows that the multi-dimensional SQT methodology is relatively robust when implemented in Central Processing Unit (CPU) environments. The section compares our method with Pitch Estimation Filter (PEF) by Gonzalez and Brookes [28] and the Normalized Correlation Function (NCF) by Atal [29]. This comparison, which was based on an independent pitch extraction test and an independent pitch labeled dataset, was conducted because PEF and NCF appeared better than three other pitch extraction techniques in the Mathworks [30] documentation: Cepstrum Pitch Determination by Noll [10], Log-Harmonic Summation by Hermes [31], and Summation of Residual Harmonics by Drugman and Alwan [32]. For this evaluation, the resolution parameters (M and N) were lowered in an early Matlab SQT implementation such that its average extraction time was slightly similar to the average extraction times of the PEF and NCF implementations that were available in the Matlab Signal Processing toolbox, as shown in Table 1. The Gaussian SQT matrix had twelve-sized spectrums ($M = 12$), but only the first five tones were used to estimate the pitch track.

The performance metrics in the comparison are thresholded Gross Pitch Error (GPE) and Root Mean

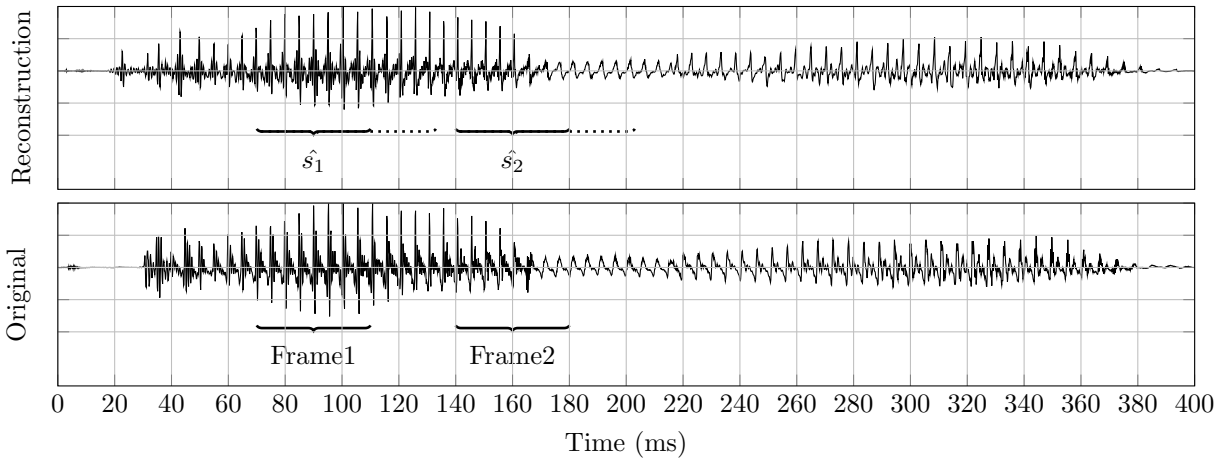
Table 1: Time Complexities of Pitch Track Extractions in Matlab

Matlab Implementations	Average Extraction Time* (%)
PEF	18.14
SQT	8.29
NCF	6.57

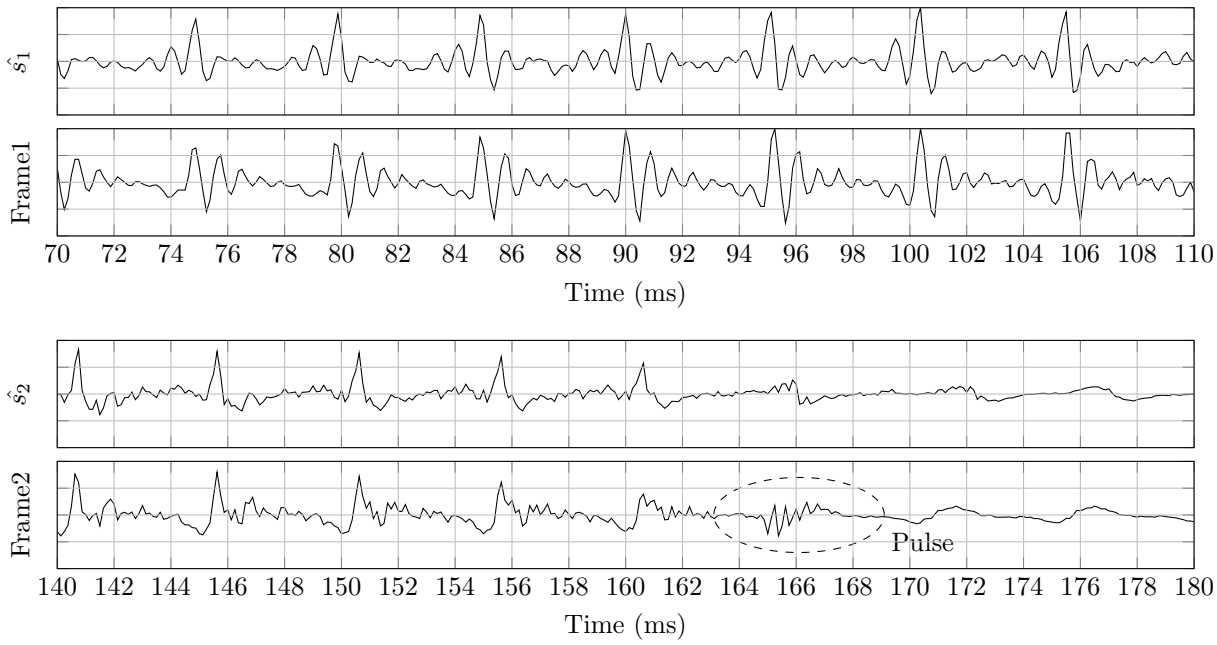
* Normalized by the average extraction time of a twelve-sized Fast Fourier Transform, which took 0.007 seconds in that machine on average.

$$\hat{s}_t[u] = \begin{cases} \sum_{m=1}^{M} b[t,u] \cdot sin(\ 2\pi \cdot p_m[t,u]\ + \pi/4\ ), & \text{if } f_m[t,u] \le 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$\left| \begin{array}{l} p_m[t,u] = \Big(\ p_m[t-1,c] - \lfloor p_m[t-1,c] \rfloor\ \Big) + \sum_{\tau=0}^{u} f_m[t,\tau] \\ f_m[t,u] = \Big(\ a[u] \cdot f_0[t] + (1-a[u]) \cdot f_0[t-1]\ \Big) \cdot m/f_s \\ b[t,u]\ \ = \Big(\ a[u] \cdot\ B[t] + (1-a[u]) \cdot\ B[t-1]\ \Big)/\sqrt{2} \\ a[u] = \min\{\ \alpha\ u/s, 1\ \}, \quad u = \ \{\ 0,1,\cdots,s-1\ \}, \quad s = \ \lfloor\ f_s/R_s\ \rfloor \end{array} \right. \tag{12}$$



(a) Reconstructed and Original Speech Functions



(b) Segment Slides

Figure 14: Reconstruction Example of the "Onward" Extracted Features

Square Error (RMSE). The thresholded GPE is a test of significance; it measures the probability that an extracted value is not within a percentage of its corresponding label value. It is used to decide whether the error of an application is acceptable or not; however, it does not show the variance of the error. On the other hand, RMSE measures the standard deviation from the error, and so it shows the scale of the error. In this case, 68.27% is the probability that error is less than the RMSE value, and 99.73% is the probability that error is less than three times the RMSE value. The smaller the divergence, the better. The metrics are defined in Equation 13. The metrics are measured under two types of common noises (white and turbine noises) and under three Signal-to-Noise-Ratio settings (20dB, 10dB, and 0dB). The background noise in the 0dB setting is on average as loud as the speech signal; 0dB is louder than 20dB. The noises were generated and mixed by a predefined Matlab function. The result of each metric was the average of 30 tests, determined at the most fitting GPE lags between the extracted pitch tracks and the target label values.

$$\begin{aligned}\text{GPE-}\xi &= P(\ |\hat{f}_0 - f_0| > f_0 \cdot \xi/100\ ) \\ &= \frac{1}{T}\sum_{t=1}^{T} x[t] \quad \Big|\ x[t] = \begin{cases} 1, & \text{if } \frac{|\hat{f}_0[t]-f_0[t]|}{f_0[t]} > \xi/100 \\ 0, & \text{otherwise} \end{cases} \\ \text{RMSE} &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left|\hat{f}_0[t] - f_0[t]\right|^2}\end{aligned} \tag{13}$$

Unfortunately, there was not an abundance of reliable labeled data when the experiments were being conducted, so the data of the verification was narrowed to a low-volume dataset known as the Fundamental Frequency Determination Algorithm (FDA) evaluation database by Bagshaw [33] although it had missing labels. The labels of the transitioning states to and from the speech activity segments were also missing in several other datasets. Therefore, the sharp edges in the pitch tracks may not have been counted in the averages. The f_0 targets of the FDA data are labeled at 20 kHz sampling rate of a 5.53-minute audio of male and female speakers. The audio was downsampled to 8 kHz before the extractions' evaluations to increase the difficulty for the pitch extractors. The outputs of the methods were also post-processed using a size-three median filter. The utilized platform was Matlab Online 2019b (9.7.0).

Figure 15 compares the GPE of the three thresholds: 5%, 10%, and 20%, where 20% is the least strict requirement or threshold of the three. The common threshold for the pitch extraction application is 10%. The figure shows that the p-value of the three methods was less than 5% when the observed value was 20%, and was less than 10% when the observed value was 10%, but the significance of the three methods was not as expected when the observed value was 5%. In either case, the figure also shows that the SQT null hypothesis is less likely than the PEF and NCF null hypotheses. The difficulty of the task was further increased by the adding additive noise. Figures 16a and 16b repeat the verification but with the white and turbine noises. The tests were repeated with the different levels of SNRs. For the 10% observed value case, the figures show that the p-value for SQT remained less than 10% under the 20 and 10dB additive noise cases. The same can be said for PEF and NCF but only under the white noise cases and the 20dB turbine environment condition noise. Finally, Figure 17 summarizes the Root Mean Square Error (RMSE) results for the seven cases. Although the performance of PEF and NCF had similar p-values, NCF appeared better than PEF by the MSE figures. Additionally, even though PEF had better MSE than NCF had in the 0dB turbine condition, the MSE error rate of NCF was obviously less affected by the white noise than PEF was.

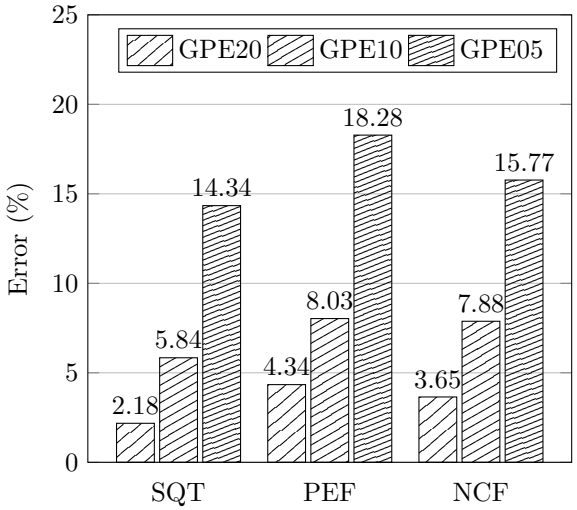
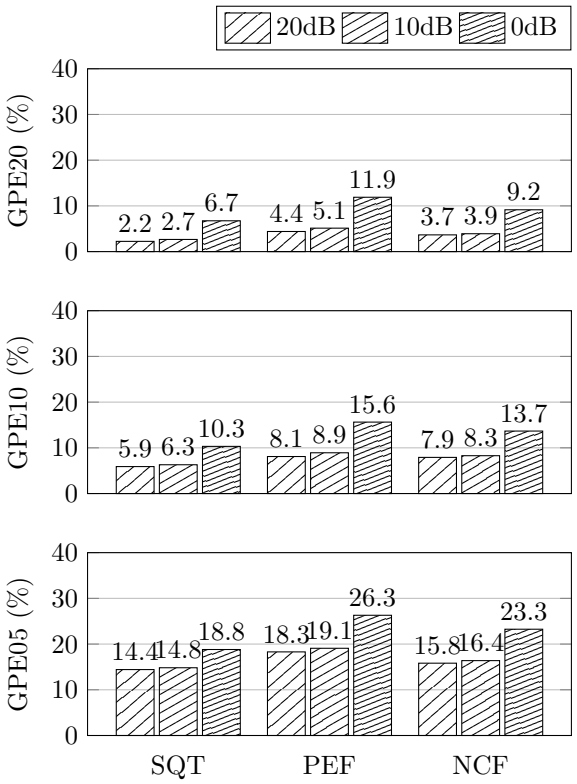
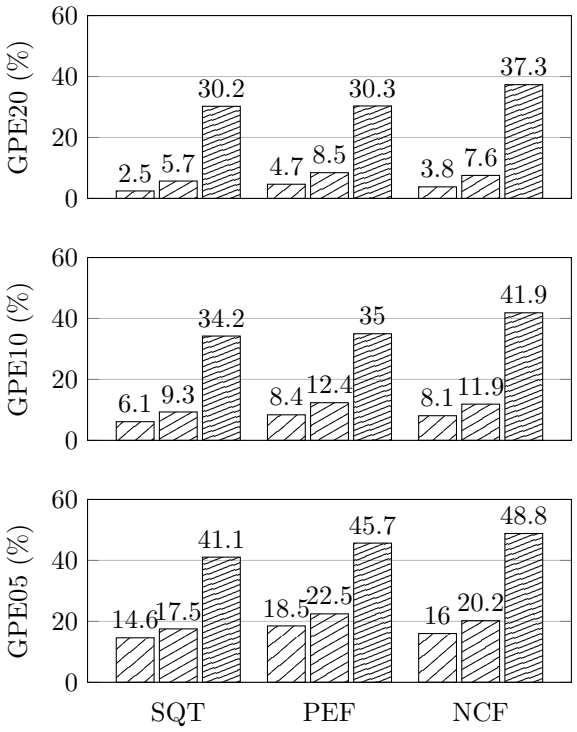


Figure 15: GPE Error Rates

In conclusion, the SQT method is as significant as the well established methods if not greater. While the SQT hypothesis appeared less wrong than 10% of the times when SNR was greater than 10dB, the PEF and NCF hypotheses appeared less wrong less than 10% of the times when SNR was greater than 20dB. The bar charts in the two figures show that white noise was easier to deal with than the turbine noise. Additionally, SQT did less error than PEF and NCF did in the RMSE cases. This was expected since the SQT error appeared correlated with the quantization error. This means the RMSE in this evaluation was a direct result of the limited resolution parameters. On the one hand, increasing the cepstral resolution (N) increases the number of the quantization levels, and so it decreases the quantization error that is due to the f_0 rounding. For example, the pitch track that is extracted using the N-16 cepstrogram in Figure 18a has a lower RMSE error than the pitch track that is extracted using the N-8 cepstrogram in Figure 18b. On the other hand, increasing the cepstral resolution (M) increases the



(a) White Noise



(b) Turbine Noise

Figure 16: GPE Error Rates, Grouped by Additive Noise Level & Type

number of observable energy components (when not capped by the sampling frequency), and so it decreases the maximum quantization value of RMS and reduces the effect of random noise on the pitch track. For example, by applying Parseval's Theorem, one can see that the detected SNR in Figure 18a is higher than the detected SNR in Figure 18b. Note that ensemble cepstrograms can be boosted gradually since the high resolution cepstrgorams can be obtained from low resolution cepstrgorams. For example, the program may extract an M-3 cepstrogram then boost M to 5 if the SNR appears low or may extract an N-8 cepstrogram then boost N to 16 if a voice activity appears in the frames, and vice versa when in energy saving mode.

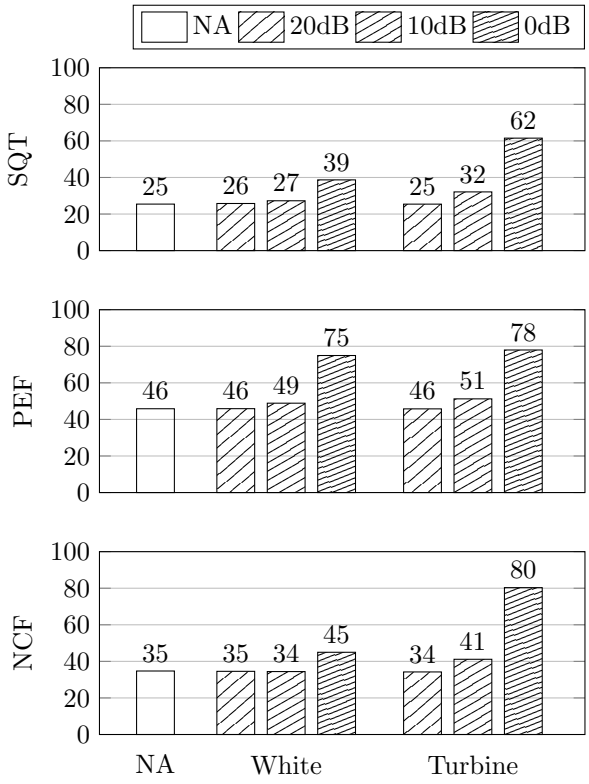
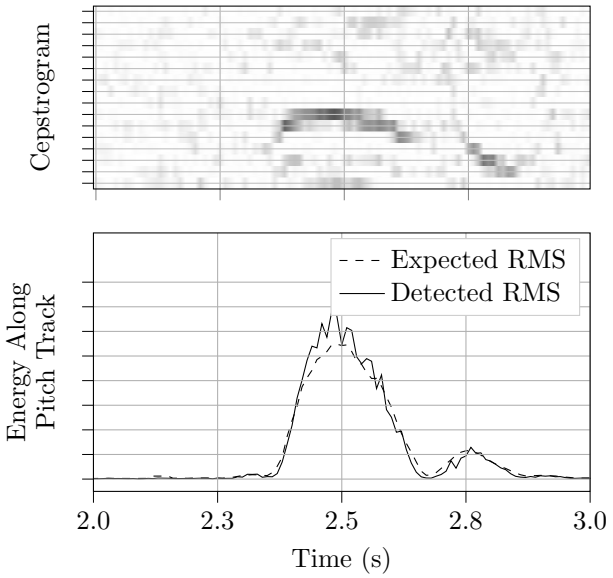


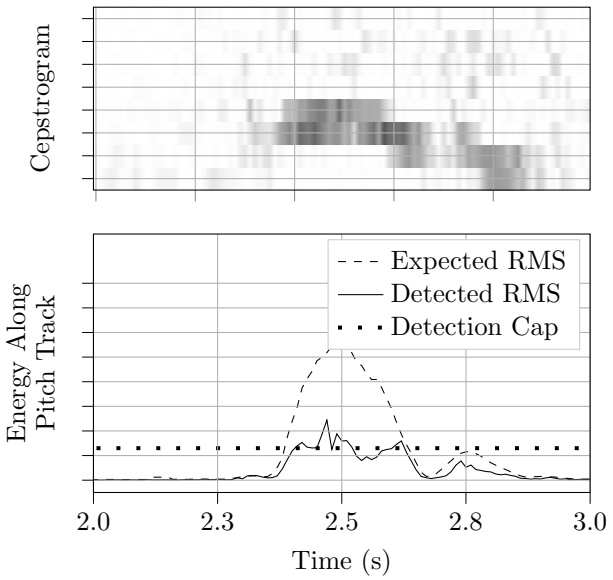
Figure 17: RMSE Error Standard Deviation, Grouped by Additive Noise Level & Type

6. Analysis

In the normal spectrogram, the formants of the high voices are more spread-out than the formants of the deep voices are. In other words, the speech that is carried in high f_0 voices occupies relatively large bandwidths, and so it may be clipped by the bandwidth channels. For example, in the 8 kHz spectrogram of Figure 20a, there are four formants of the deep voice but only three fromants of the high voice. This is because the signal was low-pass filtered before it was recorded to reduce the storage space. Similarly, the medium of the air particles is also a low-pass channel; its attenuation correlates positively with the acoustic frequency according



(a) M-5, N-16 Mode



(b) M-3, N-8 Mode

Figure 18: Adaptive Cepstral Feautres

to Islam et al. [34], Kapoor et al. [35]. However, the vast majority of the speech information is conveyed by the low frequency components that carry the first speech formant. Since the formants are the speech features that represent the different phonemes, the normalization and the extraction of the formants' features optimize both the storage and the telecommunication of the speech signals.

The normalization of the formants is achieved in the responsive spectrograms by slicing the multi-dimensional feature space to collect the harmonic energies of the pitch track. In other words, extracting the harmonic series of the fundamental frequency regulates the scale and the bandwidth of the formant features. Note that indexing is the only operator that is required for obtaining the responsive spectrogram once the multi-dimensional method is applied for obtaining the pitch track. The formants are normalized when its frequency axis is scaled responsively such that there is an alignment between the order of the formants of the similar voiced signals. For example, the pattern of the first two formants had the same scale in Figure 20 for the "Voyager" WUWs that were uttered by two different speakers. The formants that are in the spectrum that resulted from SQT (Figure 20) are more aligned than the formants that are in the corresponding fixed-band spectrum in Figure 20a.

Because the pitch track normalizes the formant features, the vocal track systems must be correlated with the glottal signals. The speech code correlates with the shape of one period of the fundamental waveform. It can be modeled in the time domain, in which it exhibits a variation of a sine cardinal function, i.e., sinc, or modeled in the frequency domain, in which it can be approximated by Gaussian Mixture Models (GMM), as shown in Figure 2. Deep voices are commonly associated with masculinity, and high voices are associated with infancy. Consequently, the gender may have a degree of correlation with the formant scaling because of the vocal folds' lengths according to Fitch [36]. However, the human voice can transit between the speech depths regardless of sex and age. Adults can produce high voices, and children can produce deep voices. For instance, parents tend to use infant-directed speech when they talk to their children. Likewise, children occasionally use adolescent-directed speech, as shown in Figure 19, which is the cepstrogram that corresponds to the spectrogram in Figure 1a.

One may speculate about the underlying physical constraints that prompt the λ_0 doubling; however, the teleport path shown in Figure 19 indicates that the f_0 has an additional dimension, which has been previously speculated in literature. For example, Hoeschele [37] and Warren et al. [38] stated a consensus of a two-dimension view which was based on experience. The pitch height, which is the frequency shift of the teleport, is believed to be the multiplicative of 110Hz, as in Equation 14. Since the f_0 is considered to be congruent to the pitch class modulo its height, the technical definition of pitch implies the angular position while the f_0 value refers to the measurement reading. Nonetheless, the height number is not constant, and the quefrency distribution can apparently be customized in several ways to increase the information gain. For example, a 2017 survey by Bernhardsson [39] in Github showed that the arithmetic population mean of the fundamental frequency varies per language. However, the uniform distribution is an optimal initiating point. Accordingly, the geometric scale, which is defined R_{Geo} in Equation 9, allocates equal quefrency pixels to the four ranges between 55, 110, 220, 440, and 880, as shown in Figure 10b. This is important because

the cepstral resolution cannot be adjusted evenly using any other scales, as in Figures 10a and Figures 10c. Adjusting the cepstrum size affects the f_0 number of levels, which do not affect the storage bandwidth as much as the spectrum size does.

$$\begin{aligned} f_0 &\equiv \text{pitch} \pmod{110} \\ &= \text{pitch} + 110 \cdot \text{depth} \quad \Big| \; \text{depth} \in \{0.5,\ 1,\ 2,\ 4,\ 8\} \end{aligned} \tag{14}$$

In terms of bit rate, the fundamental frequency can be represented by six bits when using the geometric scale by Equation 15. Additionally, the first three formants may be represented by Gaussian Mixture Model (GMM) in six variables: three means and three variances (i.e., K=3). The bit depth of the GMM variables can be ten bits per variable when using the responsive spectrogram. The bit patterns may be further compressed statistically by variable-bit-rate quantization. Nevertheless, when the frame rate (R_s) is 30 frames per second, the minimum bit rate of the speech signals is theoretically 1,980 bits per second (bps). In practice, however, any bit error in the pitch track can negatively affect the reconstruction, so its bits may have to be stored or transmitted redundantly in gray code for bit error control. The quality of the audio also heavily relies on the pitch extraction. The SQT approach is like MPEG-1 Audio Layer 3 (MP3), which was standardized in Standard of International Organization for Standardization/International Electrotechnical Commission [40] and is commonly used in the Internet for the relative quality despite the files' sizes. It is considered a lossy compression technique because it extracts only the few important frequency components from the vastly sparse frequency space. Even more efficiency may become vital as the Internet traffic and number of network nodes continue to increase. The number of connected Internet of Things (IoT) devices is expected to double within five years from the 2020 estimated figure of 11.3 billion according to Sinha [41], generating a predicted value of more than seventy Zettabytes (73.1×10^{21} bytes) of data by 2025 according to Jovanovic and Vojinovic [42].

$$\text{BitRate} = R_s \cdot (6 + 2 \cdot K \cdot \text{BitDepth}) \quad \text{(bps)} \tag{15}$$

7. Conclusion

We showed (in Section 4) how to obtain three outputs: the filtered cepstrogram, the sharp pitch track, and the reconstructable responsive spectrogram. The responsive spectrogram is sliced using the pitch track, which is extracted using the cepstrogram. The cepstrogram was derived (in Section 3), and the pitch track was evaluated (in Section 5). The reconstructability was analyzed (in Sections 5 & 6). The approach of Speech Quefrency Transform (SQT) is mainly a quefrency filter model that is achieved by spectral filtering. It transforms the speech signal directly from the time domain to the quefrency domain. In the multi-dimensional SQT method, the quefrency dimension is expanded along the frequency, phase, and sign dimensions. This hyperspace methodology appears more accurate than an alternative frequency-demodulation wavelet method although it may appear less efficient were not there speech reconstruction. The outlined procedure can be applied in everyday web applications, such as adaptive-bit-rate speech streaming; in advanced ASR deep learning, such as emotional diagnosis in medical fields; and in scarce-bandwidth telecommunication, such as deep sea and outer space exploration fields.

The SQT model was numerically using Gaussian and rectangular windows that control the spectral leakage to form the quefrency filters. While the Mel-frequency scale of MFCC does not make the cepstral features reconstructable, our multi-dimensional method adjusts the frequency banking implicitly in the frequency responses of the SQT filters. The SQT window length is defined per harmonic-frequency set since quefrency is a measurement of acceleration (cycle per square second) just as frequency is a measurement of velocity (Hertz). We applied several theories, whose effect is well known in the time and frequency domain, on the SQT quefrency domain. For example, at least two f_0 cycles must be enclosed in the SQT filters since they measure the stationarity. Because of the flexibility of the multi-dimensional approach, the logarithmic operator did not have to be included in the quefrency definition, and several quefrency scales were feasible. Even though the SQT technology outperformed several other techniques, there may still be better versions to come.

Supporting a literature hypothesis, we defined the Geometric quefrency scale (R_{Geo}) and embedded it to the novel SQT process that extracts the cepstrogram, the pitch track, and the responsive spectrogram. Although the quefrency scale can be distributed statistically using domain specific prior knowledge, R_{Geo} may be optimal for most speech applications because of its apparently

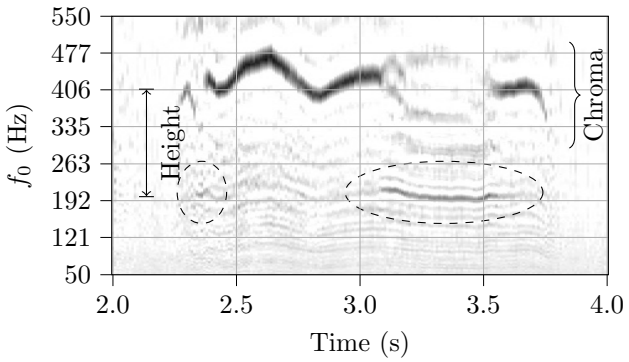
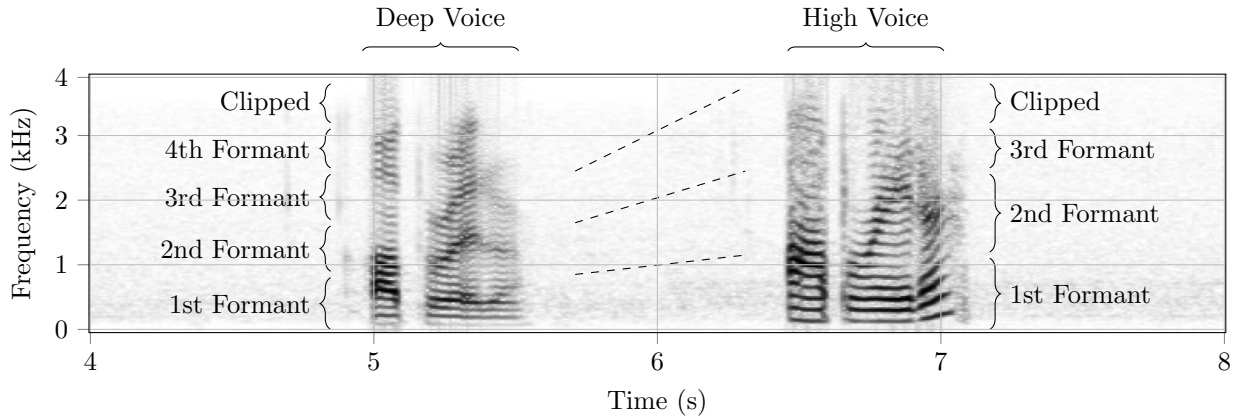
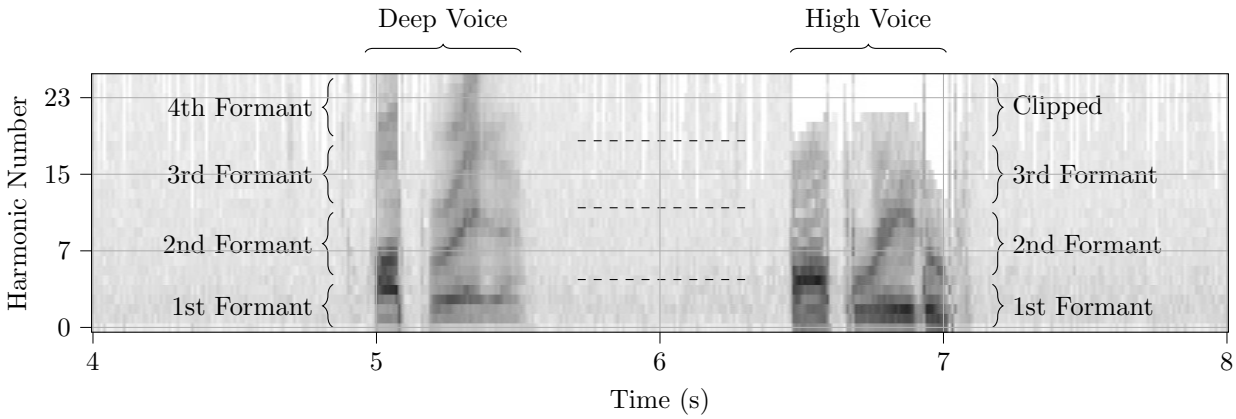


Figure 19: Cepstrogram of Human Babble. The height of the jump discontinuity of the pitch track in the graph is slightly less than 220 Hz (A_2).



(a) Fixed Band. The formants in the graph are unaligned.



(b) Responsive Band. The formants in the graph are aligned.

Figure 20: Spectrgorams with Different Formants' Alignments

equally-likely pixel probabilities. Similar to the findings of Deutsch et al. [43], our findings showed that the f_0 patterns can be expressed in two dimensions: the depth (or height) and the intonation (or pitch). The two dimensions must exist since the f_0 patterns apparently have a spectral cylindrical coordinate and since the patterns can instantly teleport between the acceleration depths. While the depth independent variable corresponds to the speech resolution because it folds up the speech code in the frequency domain, the intonation appears crucial for Wake-Up-Word (WUW) and emotion recognition. It can communicate urgency and breath patterns. It may also be beneficial for Natural Language Processing (NLP) since it can indicate syllabus stresses.

Although the size of the SQT transform in this article process is $4MN$, where M and N are the spectral and cepstral resolutions, the process appeared as robust as state-of-the-art techniques if not better. It appeared robust even in noisy conditions, and its speech reconstructions appeared viable for practical storage and telecommunication products since it pushes the speech compression rate to the lower limits. The proposed process achieved a relatively very low Root Mean Square Error (RMSE) and p-values. Its pitch tracks are estimated robustly from its fast cepstrograms that can operate on low energy for voice activity detection. The SQT cepstrograms can also be boosted due to its flexible quefrency scale. These SQT specifications are important since it is crucial that Automatic Speech Recognition (ASR) systems effortlessly detect and compose speech in resolutions that conserve energy, optimize the features' SNR, and preserve the speech components during extraction. Attuning the speech assistant to the speech of the client attenuates the background noise. The proposed process appeared facilitating Multi- and Distant-Speech Recognition.

The ASR systems with robust speech feature extractor and producer may possibly end up having a human-like learning phase as well as subjectivity and perhaps artificial feelings. This is because speech processing is apparently equivalent to Artificial Intelligence (AI) when the domain of the speech spans multiple days instead of minute sessions. The hierarchical spans of language models may simulate the intelligence levels that constitutes the

self-aware agents. In other words, the ability of comprehending logic and sequential series of events must have been, to a certain degree, built upon the primal ability of recognizing sensory data, one of which is speech. For example, increasing the contrast of sensory data may extend the attention spans according to Asiry et al. [44]. Accordingly, the resolution of the speech features can affect the word recognition quality, which can affect the grammar recognition quality. Since the recognition of temporal sequences can also enable the cognition of long-term chronological occurrences, which, when optimized, intellect possibly emerges, then the resulted SQT hyperspace may boost the virtual assistants.

Statements and Declarations

We disclose no conflicts of interest. The SQT programs are open-source. This research study is original, and the data that supports our findings can be requested from Dr Veton Kepuska, Dr Paul Bagshaw, and Minnesota Department of Health, but restrictions may apply to some of the data availability. However, the data is available from the corresponding author upon reasonable requests and permission from the original source.

References

1. Minnesota Department of Health. (2020). Data sample number 4: Baby behavior. MDH. https://youtube.com/EYe0ee2-uS4.
2. Stefanatos GA, Green GG, Ratcliff GG. (1989). Neurophysiological evidence of auditory channel anomalies in developmental dysphasia. American Medical Association: Archives of Neurology, 46(8), 871–875.
3. Bogert BP. (1963). The quefrency alanysis of time series for echoes; Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. John Wiley & Sons Inc Time Series Analysis, 209–243.
4. Rabiner L, Cheng M, Rosenberg A, McGonegal C. (1976). A comparative performance study of several pitch detection algorithms. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(5), 399–418.
5. Hess W. (2012). Pitch determination of speech signals: Algorithms and devices. Springer Science & Business Media, 3.
6. De La Cuadra P, Master AS, Sapp C. (2001). Efficient pitch detection techniques for interactive music. International Computer Music Conference (ICMC), 403–406.
7. Kawahara H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. IEEE International Conference on Acoustics, Speech, and Signal Processing, 21303–1306.
8. Charpentier F. (1986). Pitch detection using the short-term phase spectrum. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 11113–116.
9. Bogert B, Healy M, Tukey J. (1962). The quefrency analysis of time series for echos. Proc Symposium on Time Series Analysis, 209–43.
10. Noll AM. (1967). Cepstrum pitch determination. Acoustical Society of America: The journal of the acoustical society of America, 41(2), 293–309.
11. Oppenheim AV, Schafer RW. (2004). From frequency to quefrency: A history of the cepstrum. IEEE Signal Processing Magazine, 21(5), 95–106.
12. Lathi BP, Green RA. (2005). Linear systems and signals. Oxford University Press New York, 2.
13. Kepuska V. (2011). Wake-up-word speech recognition. InTech Open: Speech Technologies, 237–262.
14. Talkin D. (1995). A robust algorithm for pitch tracking (RAPT). Speech Coding and Synthesis, 495518.
15. Ewender T, Hoffmann S, Pfister B. (2009). Nearly perfect detection of continuous F_0 contour and frame classification for TTS synthesis. Tenth Annual Conference of the International Speech Communication Association, 100–103.
16. Atlas L, Janssen C. (2005). Coherent modulation spectral filtering for single-channel music source separation. IEEE International Conference on Acoustics, Speech, and Signal Processing, 4iv–461.
17. Li Q, Atlas L. (2008). Coherent modulation filtering for speech. IEEE International Conference on Acoustics, Speech and Signal Processing, 4481–4484.
18. Ganchev T, Fakotakis N, Kokkinakis G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. Proceedings of the International Conference on Speech and Computer (SPECOM), 1191–194.
19. Moorer J. (1974). The optimum comb method of pitch period analysis of continuous digitized speech. IEEE Transactions on Acoustics, Speech, and Signal Processing, 22(5), 330–338.
20. Zahorian SA, Hu H. (2008). A spectral/temporal method for robust fundamental frequency tracking. The Journal of the Acoustical Society of America (ASA), 123(6), 4559–4571.
21. Kawahara H, Masuda-Katsuse I, De Cheveigne A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Elsevier: Speech Communication, 27(3-4), 187–207.
22. Ykhlef F, Ykhlef H, Aissat A. (2012). Influence of Dolph-Chebyshev window on speech enhancement. IEEE International Conference on Multimedia Computing and Systems, 140–143.
23. Wölfel M, McDonough JW. (2009). Distant speech recognition. Wiley Online Library.
24. Rabiner LR, Gold B. (1975). Theory and application of digital signal processing. Prentice-Hall Inc, Englewood

Cliffs, NJ, 777.
25. Standard of International Organization for Standardization. (2003). Acoustics: Normal equal-loudness-level contours. ISO 226:2003, Geneva CH.
26. Mehrabian A. (1972). Nonverbal communication. Transaction Publishers.
27. Green JA, Whitney PG, Potegal M. (2011). Screaming, yelling, whining, and crying: Categorical and intensity differences in vocal expressions of anger and sadness in children's tantrums. American Psychological Association: Emotion, 11(5), 1124.
28. Gonzalez S, Brookes M. (2011). A pitch estimation filter robust to high levels of noise (PEFAC). IEEE 19th European Signal Processing Conference, 451–455.
29. Atal BS. (1972). Automatic speaker recognition based on pitch contours. The Journal of the Acoustical Society of America, 52(6B), 1687–1697.
30. Mathworks. (2019). Estimate fundamental frequency of audio signal. Matlab Docs R2019b. https://www.mathworks.com/help/releases/R2019b/audio/ref/pitch.html.
31. Hermes DJ. (1988). Measurement of pitch by subharmonic summation. The journal of the Acoustical Society of America, 83(1), 257–264.
32. Drugman T, Alwan A. (2019). Joint robust voicing detection and pitch estimation based on residual harmonics. arXiv Preprint 200100459.
33. Bagshaw P. (1993). Fundamental frequency determination algorithm (FDA) evaluation database. University of Edinburgh's Centre for Speech Technology Research.
34. Islam MR, Elshaikh ZEO, Khalifa OO, Alam AZ, Khan S, Naji A. (2010). Prediction of signal attenuation due to duststorms using Mie scattering. IIUM Engineering Journal, 11(1), 71–87.
35. Kapoor R, Ramasamy S, Gardi A, Schyndel RV, Sabatini R. (2018). Acoustic sensors for air and surface navigation applications. MDPI Sensors, 18(2), 499.
36. Fitch WT. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. The Journal of the Acoustical Society of America, 102(2), 1213–1222.
37. Hoeschele M. (2017). Animal pitch perception: Melodies and harmonies. Europe PMC Funders' Comparative Cognition & Behavior Reviews, 125.
38. Warren J, Uppenkamp S, Patterson RD, Griffiths TD. (2003). Separating pitch chroma and pitch height in the human brain. Proceedings of the National Academy of Sciences, 100(17), 10038–10042.
39. Bernhardsson E. (2017). Lang-pitch. GitHub. https://github.com/erikbern/lang-pitch.
40. Standard of International Organization for Standardization/International Electrotechnical Commission. (1993). Information technology: Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s part 3: Audio. ISO/IEC 11172-3:1993, Geneva CH.
41. Sinha S. (2021). State of IoT 2021: Number of connected IoT devices growing 9% to 12.3 billion globally, cellular IoT now surpassing 2 billion. IoT Analytics. https://iot-analytics.com/number-connected-iot-devices.
42. Jovanovic B, Vojinovic I. (2021). 45 fascinating IoT statistics for 2021: The state of the industry. Data Prot Net. https://dataprot.net/statistics/iot-statistics.
43. Deutsch D, Dooley K, Henthorn T. (2008). Pitch circularity from tones comprising full harmonic series. The Journal of the Acoustical Society of America, 124(1), 589–597.
44. Asiry O, Shen H, Wyeld T, Balkhy S. (2018). Extending attention span for children ADHD using an attentive visual interface. IEEE 22nd International Conference Information Visualisation (IV), 188–193.