

Speech Quefrequency Transform (SQT)

Ahmad Zuhair Hasanain (✉ hasanainaz@gmail.com)

Umm Al-Qura University <https://orcid.org/0000-0003-0695-3657>

Veton Z Kepuska

Florida Institute of Technology

Marius Calin Silaghi

Florida Institute of Technology

Ivica N Kostanic

Florida Institute of Technology

Research Article

Keywords: Speech Extractions, Pitch Tracks, Enhanced Cepstrograms, Normalized Spectrograms, Speaker Normalizations, Speech Reconstructions

Posted Date: July 18th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-779995/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Speech Quefrency Transform (SQT)

A Robust Speech Features Extractor & Reconstructor For Artificial Intelligence Frontends

Author1 · Author2 · Reviewer1 · Reviewer2

Received: date / Accepted: date

Abstract Human speech consists mainly of three components: a glottal signal, a vocal tract response, and a harmonic shift. The three respectively correlate with the intonation (pitch), the formants (timbre), and the speech resolution (depth). Adding the intonation of the Fundamental Frequency (FF) to Automatic Speech Recognition (ASR) systems is necessary. First, the intonation conveys a primitive para-language. Second, its speaker-tuning reduces background noises to clarify acoustic observations. Third, extracting the speech features is more efficient when they are computed together at the same time.

This work introduces a frequency-modulation model, a novel quefrency-based speech features' extraction that is named Speech Quefrency Transform (SQT), and its proper quefrency scaling and transformation function. The cepstrums, which are spectrums of spectrums, are suggested in time unit accelerations, whereby the discrete variable, the quefrency, is measured in Hertz-per-microsecond. The extracted features are comparable to Mel-Frequency Cepstral Coefficients (MFCC) integrated within a quefrency-based pitch tracker. The SQT transform directly expands time samples of stationary signals (i.e., speech) to a higher dimensional space, which can help generative Artificial Neural Networks (ANNs) in unsupervised Machine Learning and Natural Language Processing (NLP) tasks. The proposed methodologies, which are a scalable solution that is compatible with dynamic and parallel programming for refined speech and cepstral analysis, can robustly estimate the features after applying a matrix

Author1
Affiliation1
Email1
Orcid1

Author2
Affiliation2
Email2
Orcid2

multiplication in less than a hundred sub-bands, preserving precious computational resources.

Keywords Speech Extractions · Pitch Tracks · Enhanced Cepstrograms · Normalized Spectrograms · Speaker Normalizations · Speech Reconstructions

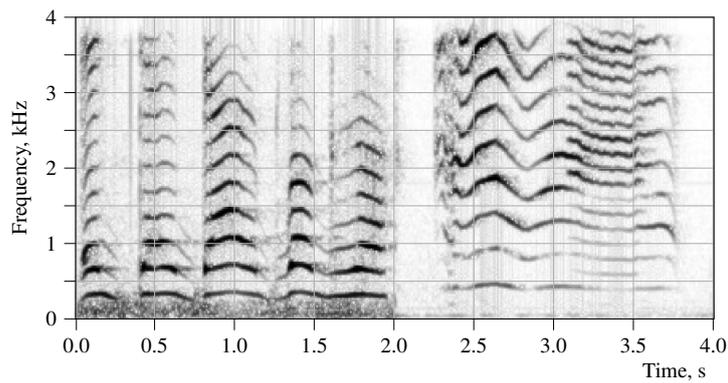
1 Introduction

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 chirrup (Figure 1b). The salient curly harmonics of the voice render a hidden state that appears contentious when connecting the dots. The speech spectrogram is a graph of the energy distribution along audio frequencies (i.e., the spectrum) versus time. The contrasts of the pictured graphs were adjusted per the CMYK printing norms; the higher the energies, the darker the pixels, but the color scheme can be reversed when printed on monitors. The human speech can be captured from the spectrogram using two features: the pitch and the harmonic intensities. In order to have artificial agents processing (or understanding) spoken languages naturally, it is crucial to realize a mathematical representation for speech that is attuned accordingly, especially because human intelligence and language are tangled up during development.

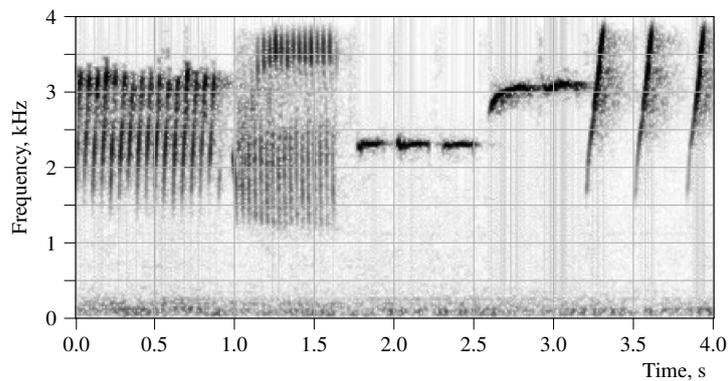
The linearly-spaced curves in the spectrogram are the outcome of a periodic time signal that is locally stationary. This means that the periodicity and the waveform shape fluctuate slowly, relative to the sampling rate. The more the signal is locally stationary, the sharper the curves are. Speech producers flap in response to internal air pressure, air molecules are compressed and released periodically, and the pulse shape makes speech transmittable through air. The time distance between two adjacent compressions (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 (FF or f_0). Per context, it is also the pitch and the frequency carrier. However, being in an air channel as its communication medium, the signal's actual periodicity is in meters per cycle. The λ_0 and ν in the equation are the corresponding wavelength and the speed of sound (in the medium). All variables are time variants. The ν is usually assumed to be constant although the temperature, humidity, and wind speed are not so along the air travel paths from the speech producer (vocal folds, cords, or glottis) to other human receivers.

$$T_0 = \lambda_0/\nu = 1/f_0 \text{ (second-per-cycle)} \quad (1)$$

In Figure 1b, the birdsong producer constrains its f_0 in a time-variant coordinate, which is then 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 (e.g., 1kHz and 4kHz) are identical, because there are unaccounted independent axes. For example, the fundamental waveform of the canary bird is visually rotated around a variable axis parallel to the time axis, and its perimeter path renders a visual effect of cylinders in the figure. Assuming the bird's mono f_0 was



(a) Child Babble



(b) Canary Twitter

Fig. 1 Spectrograms of Multi- and Mono-Resonance Communication Systems

traveling with a constant angular velocity in a polar coordinate, the inferred radius of a pictured time-variant cylinder is about 1kHz and centered at 2kHz.

Similarly, the infancy voice in Figure 1a is teleported back and forth between two speech depths, which can be noticed in the audio playback, during moments of emotional outbursts (e.g., between 3.25 and 3.6 seconds). Per the juxtaposition of the two spectrograms, the human voice had a fundamental waveform, whose shape was transforming at a slow pace and had harmonic components, which rendered the parallel spectral curves. The tone-change phenomenon usually happens during puberty, and it doubles the fundamental interval, gears down the speech depth, and folds up the spectral code bandwidth. This creates the speech resolutions. It is commonly known that deep human voices have been relatively overrepresented in the classic telephony bandwidth (4kHz). It is also known that the spectral bandwidth scaling has been one of the main challenges in Automatic Speech Recognition (ASR). In order to equalize the speech features extractions, more coordinates must be added up.

The spectral energies of the harmonic components is the speech features. If the parallel curves were the openings of window blinds, a few shaded patterns would appear behind the blinds. These patterns are called speech formants, and their mixtures make multitone phonemes. The harmonic components are also called timbre, overtones series, and cepstral coefficients. In this work, they are also the frequency envelope, the frequency-modulating signal, and the vocal tract response (H_m) for the purpose of mathematically modeling the speech signals and systems. 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.

A 1989 study Stefanatos et al. (1989) suggested that human perception of speech relates to frequency demodulation. It also happens that two conventional approaches to pitch and speech feature extractions are modulation-based but applied to the frequency domain. Unfortunately, the classical frequency domain, albeit vital, made several speech modules computationally expensive. The speech modules that are needed are namely: fundamental frequency acquisition and tracking, spectral depth normalization, and speech signal generation. Processing natural languages, for instance, the Generative Adversarial Networks (GANs), needs a robust acoustic frontend that effortlessly unpacks and composes speech utterances.

This work presents a novel modulation-based methodology for extracting the fundamental frequency and the harmonic energies, consisting of a speech model, a transform, and features extraction methods. The model is based on axiomatic assumptions: The voiced utterances are expressed in a variable speech resolution, and extracting the piloting f_0 makes the speech code instantly attainable. Per the model, the human voice utilizes 25.0 kilo-bit-per-second (kbps) and is intelligibly recoverable when confined to a 4.1 kbps bandwidth. An overview, related work, and the contributions of this work are in Section 2. Sections 3, 4, and 5 outline the approach, the methodology, and the analysis, respectively. The approach embraces the quefrequency scale, the signal model, and the cepstral measure. The methodology describes the quefrequency transform and the speech feature extraction. The analysis highlights the findings with discussion.

2 Review

The pitch extraction techniques are generally categorized into temporal, spectral, and cepstral approaches based on the processing domain: time-lag, frequency, and quefrequency. The quefrequency is the frequency of frequencies and is the independent variable of the cepstrum, like the frequency is so of the spectrum. Even though several techniques had existed for pitch extraction Rabiner et al. (1976), it was still one of the most computationally demanding modules Chen and Hu (2007).

One of the time-domain methodologies is auto-correlation, which matches the speech signal with its lagged versions (as opposed to decomposing its independent frequency components). The frequency-domain is obtained in three operations: windowing, Fast Fourier Transform (FFT), and frequency banks. It is noteworthy that

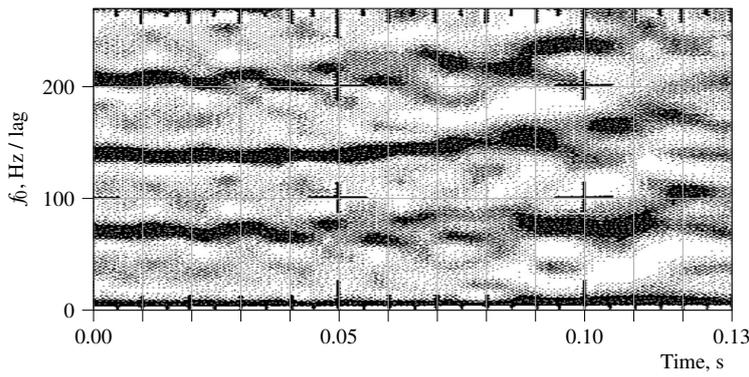
although the inverse Fast Fourier transform (iFFT) gives exact inversion theoretically, a Fourier transform defined over a finite interval is actually non-invertible, so its iFFT inverse is an estimation.

Like the frequency domain, the quefrency domain is a non-linear vector space, in which superposition does not hold. In the cepstrogram, the signals of the tract and the glottis become separable due to their characteristic differences. The cepstrogram is usually obtained from a high-resolution spectrogram along with two additional operations: logarithmic scaling (on both magnitude and frequency) and an iFFT. The iFFT is further approximated using the Discrete Cosine Transform (DCT) in the Mel-Frequency Cepstral Coefficients (MFCC) method. The cepstrum analysis Bogert et al. (1962) is defined as the power spectrum of the log magnitude of the power spectrum of the time samples Oppenheim and Schaffer (2004); i.e., Equation 2 or, equivalently, $\mathcal{F}^{-1} \log |\mathcal{F}\{\bullet\}|$, where $\mathcal{F}\{\bullet\}$ denotes a forward Fourier transformation, $e^{-j\theta} = \cos(\theta) - j \cdot \sin(\theta)$ Rabiner and Schaffer (2011); Lathi and Green (2005), and $j = \sqrt{-1}$. In this work, the SQT domain emulates the set of the fundamental waveforms, which is a subset of periodic functions.

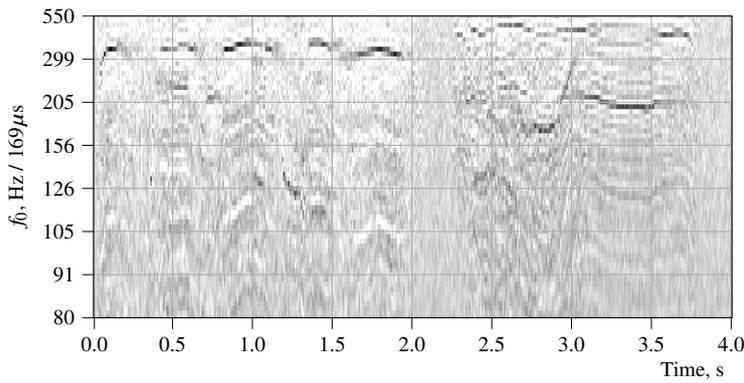
$$p[n] = \frac{1}{2c+1} \sum_{m=0}^{2c} e^{j2\pi \frac{nm}{2c+1}} \log \left| \sum_{u=0}^{2c} e^{-j2\pi \frac{mu}{2c+1}} s[u] \right| \quad (2)$$

For the pitch extraction, a common method is the harmonic product spectrum, which De La Cuadra et al. (2001) showed effective in adjusting acoustic instruments. Moreover, several methodologies in literature appeared promising, such as Pitch Contours (PC) Atal (1972), Amplitude Compression Gonzalez and Brookes (2011) (AC), and neighbor normalization Talkin (1995). However, David Talkin commented on the latter that its Normalized Cross-Correlation (NCC) amplified the "peak at twice the correct period," referring to what is called an "A" tone. Figure 2a Talkin (1995) shows three possible pitch tracks at 70Hz, 140Hz, or 210Hz. According to the source, an NCC-based method was not instantaneously able to differentiate between the pitch and the two harmonically similar tones, which are apart at a one-unit-octave frequency. The "almost perfect pitch tracker" of Ewender et al. (2009) showed the challenge of the overtones has still occurred in this century literature.

Another recent approach to speech features was the coherent modulation Atlas and Janssen (2005); this method is known to be "useful" when the modulating and modulated signals have commonalities. The coherent modulation was "more effective than previously believed," stated Atlas and Clark Clark and Atlas (2009). One of Atlas's pieces of work Li and Atlas (2008) outlined a detailed FFT-based procedure. Generally, the bulk of the algorithms in the literature perplexes with the illusionary overtones because of the f_0 harmonic characteristic. Even human perception regards them more similar than other tones, but human agents easily differentiate between each speech depth, implying that the ambiguity in f_0 should not exist in the acoustic frontend, rather at a deeper perceptual level. For instance, one could argue that harmonic illusion may have played a role in easing the language acquisition since it rolls the spectral code of the main speakers' categories.



(a) NCC-Based Talkin (1995)



(b) SQT-Based

Fig. 2 Cepstrograms of Two Speech Feature Extractors

On the other hand, the frequency domain approaches are mediocre at limited computational power, and customizing the conventional frequency axis requires frequency banks. According to Moorer Moorer (1974), the temporal and cepstral methods can complement one another as their features are approached differently. Unsurprisingly, the cutting-edge pitch trackers, such as Yet Another Algorithm for Pitch Tracking (YAAPT) Zahorian and Hu (2008) of Zahorian and Hu, combine observation candidates from more than one pitch detection approach before smoothing the estimated pitch track. This combination comes with a computational cost, at which, the sole cepstral approach is capable of producing fine pitch tracks. A possible workaround to this issue is to apply the process adaptively. Adaptive methodologies vary in the literature. For example, Kawahara et al. Kawahara et al. (1999) managed to suppress the overtone noise but with a smoothing operator that flattened the pitch track and, consequently, reduced the accuracy of the f_0 readings. The smoothing is an averaging operator that distorts acoustic information. It has to be replaced with a non-linear filtering to preserve the original information recoverability. Several cutting-edge approaches have

not been as robust as mammals, whose biological processors demodulate and compose speech signals effortlessly even though they operate at much slower processing speeds than today's machines.

SQT is novel and distinctive. Unlike the previous methodologies, not only does SQT extract recoverable harmonic spectra, but it also distinguishes the f_0 from its over- and undertones during the pitch extraction phase. Figure 2b depicts the cepstrogram of a multi-speaker channel. Before conducting the SQT-based extractions, male and female voices were added to an infancy voice background. The two persons had their f_0 in the quefrency ranges [80, 155)Hz and [155, 250)Hz, respectively, while the infant had its in the range [250, 550]Hz. Theoretically, the addition of periodic signals is irreversible. However, each of the male and female utterances was still intelligible when "unmasked" or recovered from the other two voices in the background. This example illustrates the advantageous noise resilience of the proposed method. SQT has unique feature spaces, responsive normalizations, and quefrency scales. SQT extracts the speech features in practical means that can satisfy the required accuracy of any application. This is because applying Nyquist's theorem, which is usually applied on time axes, on the frequency axis suppresses the quefrency aliases. The outcome of this is analogous to the human ability to tune to one speaker and having the ambient noises blurred. In other words, SQT reduces the resolution of the background speeches. The other cepstral, spectral, and temporal approaches were either ineffective or computationally prohibitive when standing alone. They postpone complexities for later post-processing, since finding the speech model is not an easy task for unsupervised Machine Learning. In contrast, SQT facilitates relatively advanced capabilities for the artificial speech agents, such as simultaneously processing multiple and distant speeches.

3 Approach

Since the human acoustic sense is receptive to frequency-modulated tones, and because the vocal features are separable in the quefrency domain, it made sense to investigate the cepstral approach. Some may argue that not all human speech is periodic because there are phonemes that are unvoiced. However, the unvoiced phonemes still have a spectral presence and are partially detected by periodic filters. Additionally, the unvoiced phonemes are not entirely unvoiced. They are usually coupled by voiced segments to increase their air transmissivity. The unvoiced units are variations of noise, such as the violet noise, and can be modeled by smaller filters since they have smaller frequency resolution. Instead of adjusting two frame rates, another way to increase the unvoiced presence is to have the frame step no larger than 10ms, or $10m f_s$ samples, where f_s is the sampling rate. The frame step is the time interval between similar points at two adjacent frames; it is also the complement of the frame overlapping. In other words, the sampled spectrograms can partially capture speech pulses and noises when the frame rate is increased due to its fast-paced transition. Note that, once the signal is in frames, the speech sequence is re-sampled from the sampling rate to the frame rate, $f_r = 1/\text{Frame_Step}$ or 100fps (frame-per-second).

The frequency domain is obtained from the quefrequency selection on the SQT domain, which is obtained directly from the time frames.

This section presents logical intuitions and infers presumption axioms for the appropriate quefrequency scale, the speech signal model, and the periodical cepstral measurement. The discussion also extends to the maximum window size.

3.1 Reciprocal Quefrequency Scale

Cepstrum is a measurement of an acceleration rate, and its quefrequency is equivalent to the change in frequency (Hz) per a time interval (μs). It may be misleading to count its unit in either seconds or samples although it had been argued (as in (Jurafsky and Martin, 2009, p. 301)). The proof that quefrequency is the rate of change can be derived directly from intuitive definitions. Let the scalar quantities λ_0 , $second$, and $second'$ be sample measurements of the corresponding units cycle, s , and s' , in some ϖ time unit. Since quefrequency is the frequency of frequencies, but frequency is the rate of cycle per the standardized second; hence, the rate of occurrence with respect to time, and since frequency also is, intuitively, the ratio of the standardized $second$ to the comparable λ_0 interval; hence, $\frac{second}{\lambda_0 \varpi}$ (in cycle/second or Hz unit), then similarly, quefrequency is $\frac{second' \varpi}{second \varpi / \lambda_0}$ (in Hz-per- s' unit). That is, quefrequency is the rate of (wave-period per the standardized second) per another constant $second'$; and therefore, the unit of the quefrequency is cycle per second squared if $second' = second$. In other words, although some equal quantities may be divided, their units are generally multiplied. For that reason, our reciprocal definition of the appropriate quefrequency scale is Equation 3, where f_{min} and f_{max} are the desired minimum and maximum quefrequencies or f_0 boundaries, and $n \in \{0, 1, \dots, N-1, N\}$.

$$R_{(n)} = R(n, f_{min}, f_{max}) = \frac{1}{\frac{1-n/N}{f_{min}} + \frac{n/N}{f_{max}}} \quad (\text{Hz}/s') \quad (3)$$

Consequently, the quefrequency spacing is $1/\Delta q = \frac{f_{max} \cdot f_{min} \cdot N \times 10^{-6}}{(f_{max} - f_{min})}$ Hertz-per-microsecond (or Hz/ μs), and most importantly, since

$$\frac{1}{R_n} (s'/\text{Hz}) = \frac{1}{R_{n+1}} (s'/\text{Hz}) + \frac{1}{\Delta q} (\mu s/\text{Hz}) \quad (4)$$

then, the equivalent temporal unit is

$$s' = \frac{(f_{max} - f_{min}) \cdot 10^6}{f_{max} \cdot f_{min} \cdot N} (\mu s) \quad (5)$$

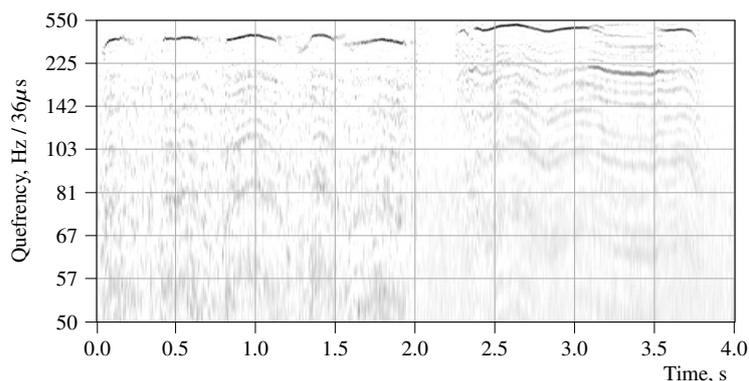
Notice the quefrequency unit factor s' and the desired quefrequency resolution N are inversely proportional. Precisely, $s' = (f_{max} - f_{min}) \cdot 343 \times 10^2 / (f_{max} \cdot f_{min} \cdot N)$ centimeters when the speed of sound is 343 meters-per-second. The unit of the quefrequency can be converted to samples given the sampling rate, as has been referred to. Notice that the Bark- and the Mel-scale, as well as the quefrequency scale, exhibit logarithmic curves. One could hypothesize that the human reception of voice has a

lower boundary (about 20Hz/s') because of the reciprocal proportion, which spans quickly to infinity for the Direct Current (DC) measurement. Figure 3a shows the practical Reciprocal Scale (R-Scale), where the quefrency bins are spaced consistently such that $\Delta q = \text{Hz}/36\mu\text{s}$. On the other hand, Figure 3b shows the regular linear scale, where Δq varies from $\text{Hz}/392\mu\text{s}$ to $\text{Hz}/3\mu\text{s}$. The two figures show cepstrograms of the same child voices, whose pitches are in [190, 480]Hz. The number of calculations was set constant to generate the two depictions, yet the linear scale is inefficient for several reasons. First, the frequencies around 100Hz/s' of the regular cepstrum received exponentially low intensity, so they might underflow without a steep increase in computational resources. Second, the width of the pitch track and the energy distribution were not equal, and applying a post-processing process, such as frequency banks, on the already enlarged side of the scale, exacerbates the complexity. Third, normalizing the frequency scale after the fact adds uncontrollable noise to the pitch track because of the digitization happening at the intermediate phases. The common spectral approaches are needing more calculations to produce an imbalanced scale that requires the presence of another complexity that directly affects the readings' accuracy. On the other hand, placing the normalization up front equalizes the precision and boosts efficiency. See Figure 3a; the bandwidth (the line thickness) of the pitch track is consistent in the proposed R-Scale, and also obtaining the cepstral range from [150, 500]Hz/s' consumed less than half the resources that would be necessary with the linear spacing. The quefrency ranges [110, 220]Hz and [220, 440]Hz get the same resolution when the two scales are weighted 70:30. The SQT scale is customizable per application.

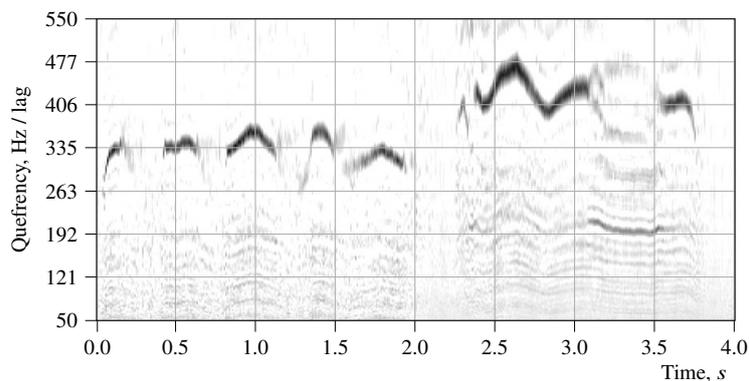
3.2 Window Size

The interval of the sliding frame corresponds to the lower limit of the fundamental frequency since f_{min} is reciprocal to the maximal length of the wave-period. Notice Equation 6 has m carrier signals (or frequency samples) that share the same phase and each is applied on a corresponding window $w_m[u]$ in the general case, putting the spectral leakage into consideration. When the quefrency scale is applied, $w_m[u] = w[u]$ for equal magnitude scaling; all harmonic elements of an f_0 are measured in the same window. Moreover, from the demodulation perspective, the quefrency sampling rate q_s has to be at least double the frequency of the modulated signal. That is, based on Nyquist's, at least two observations per cycle are necessary for a reliable detection. The minimal sampling rate q_s required to check if $\cos(2\pi t \cdot f_0)$ exists without aliasing is $\frac{2}{f_0}$. Accordingly, the minimal length of the window is $(2c + 1) = \frac{2}{f_{min}}$ seconds; equivalently, $c \geq \lceil f_s / f_{min} - 0.5 \rceil$ samples, where the $\lceil \bullet \rceil$ ceils \bullet up to the nearest integer.

There are three aspects to consider in determining the minimal quefrency. According to psychoacoustic experiments, human hearing may sense frequencies down to 20Hz and does not discriminate between acoustic echos lagging less than 100 milliseconds Wölfel and McDonough (2009). Rarely does a speech signal have a frequency of less than 60Hz, and these constraining values increase as one advances in age. For most people, however, human perception is most responsive to the frequency



(a) Reciprocal



(b) Linear

Fig. 3 Quefrency Scales

band [1, 5]kHz. Having said that, one should set the lowest measurable fundamental of the speech model to $f_{min} = 50\text{Hz}$. This value sets the minimal frequency required to 25Hz (this is approximately the lowest perceivable frequency to human beings) and the window length to 40.125ms (less than the length of the ambiguity). Moreover, the spectral range of the audio instruments is above 50Hz (Rabiner and Gold, 1975, p. 140). The silence at 50Hz corresponds to 40dB, at the lowest contour of loudness per the standardized Equal-Loudness Contours (ISO 226:2003) and according to the 1956 work of D. Robinson and R. Dadson. In summary, for 8kHz sampling rate, 321 subsequent samples are required in order to detect an event (e.g., impulse) that happens every 161 samples, and the shortest stationary unit of spoken languages takes more than 40 milliseconds.

3.3 Signal Model

Because speech is sampled, its harmonic sequence is bounded. The continuous speech signal is a random process whose random variables are the vocal tract and folds states. The two have physical limits safeguarding their characteristics from abrupt changes. This makes speech presentable in frames. Let each time frame have $(2c + 1)$ subsequent samples such that the discrete time axis is $u \in \{0, 1, \dots, 2c\} = \mathbb{Z}_{[0, 2c]}$. Let a harmonic number (also called order, rank, or term) be $m \in \{1, 2, \dots, M\}$, where M is the highest desired harmonic order. In addition, let a set of fundamental frequencies have indices $n \in \{0, 1, \dots, N\}$. The variable lengths N , M , and c are non-negative integers (\mathbb{Z}_+) and are utilized in Section 4.2. To obtain the signal model, it is of interest to formulate a quefrency transform $T : \mathbb{R}_{[-1, +1]}^{(2c+1)} \rightarrow (\mathbb{R}_{[f_{min}, f_{max}]}^{(N+1)}, \mathbb{R}_+^M)$ that maps the $(2c + 1)$ -sample vector to an $(N + 1)$ -quefrency \times M -harmonic matrix.

Without loss of generality, consider the situation where only one glottal signal and one vocal tract system generate the time frame. The speech is a time variant whose waveform can be modeled with wavelets. Needless to say, the harmonic multiplicities are embedded within the f_0 waveform, which is a variable function. However, in the ideal case, the glottal signal is a unit impulse train, whose Fourier transform is a train of impulses, re-scaled to $2\pi f_0$ magnitude (noted later in Equation 10). The idealization of the basic waveform $g[u]$ spreads its f_0 into harmonics with equal magnitudes. Now, placing four formants on that harmonic medium can be obtained when the impulse train passes through a multi-bandpass filter, whose impulse response function h is actually the shape of the waveform. The convolution happens as the filtering tract system responds with a time sequence at every stimulus it receives. Moreover, due to the speech depth phenomenon, the filtering system (h) is a nonlinear function of f_0 . In general, a speech time frame is modeled as Equation 6, where the summation is over the harmonics, the convolution ($*$) is the filter operator, and the φ is the instantaneous phase, usually in the range $[0, 2\pi]$ radius. However, since the phase is considered only for synchronization (minimizing the angular difference between the signal transmission and the filter reception), and because the output of interest is the absolute magnitude, its effective range becomes $[0, \frac{\pi}{2}]$ radius, and an absolute value operator is added for the other half. The function $\varphi[\omega]$ for d -index phase $\omega = \{0, 1, \dots, d - 1\}$ is defined in Equation 7 for completion.

$$s[u] = \underbrace{\sum_m w_m[u, f_0] \cdot \cos(2\pi \frac{u}{f_s} \cdot m f_0 - \varphi)}_{g[u]} * h[u, f_0] \quad (6)$$

$$\varphi[\omega] = \frac{\pi\omega}{2(d-1)} \quad (7)$$

A practical enhancement to reduce the spectral leakage is to define the time-frame windowing $w[u, f_0]$, controlling the widths of the harmonic banks because the default window is square if the shape is not defined. The frequency response of the window substitutes each sinusoid filter with a continuous range of sinusoids, resulting in a relatively wider band at $m f_0$. The windowing also attenuates the energy

magnitudes exponentially. The Dolph-Chebyshev function is one of the unique windows because it has an almost flat spectral attenuation. In other words, the additive distortion of the Chebyshev window is roughly distributed uniformly, enhancing the performance of a subsequent maxima detection. Frame-wise, the windowing operator is linear and reversible as long as its entries are positives. The Chebyshev window is achieved per the desired width of the main lobe and an iFFT application. Applying the Stone-Weierstrass theorem, the windowing generates polynomial functions that realize approximated cosine functions on the bounded bandwidth. Keep in mind that the equality of Parseval's theorem is defined on rectangular windows.

Given the equation of the main-lobe width defined in Smith (2011 edition), one may obtain Equation 8, which calculates the level of attenuation (A) in decibels (dB), having a frame length $(2c + 2)$ and a main-lobe side width w_b . From the cepstral perspective, the spectrum of $h[u, f_0]$ ($H_m[f, f_s]$) is sampled at $q_s = \frac{1}{f_0} \leq \frac{1}{f_{max}}$ (in harmonic-per-Hz). Crystallizing the definition of the quefrency unit of the previous section, the application of Nyquist's theorem on the quefrency domain makes the spectral width less than or equal to $2f_0$; i.e., $2w_b \leq 2/q_s$. The desired width is $w_b = f_0/|2\kappa - \sigma|$, where $\sigma \in [0.5, 1.5]$. When $\sigma = 0.5$, the model is resilient to noise, and when $\sigma = 1.5$, the model is enhanced in noise-free environments. Since most Chebyshev implementations require the attenuation, the included equation is handy. The default value of σ is 1.0. Equation 9 defines a Gaussian-based alternative window. Most importantly, estimating the energy after the transform is possible only when the samples of the window sum to one. Finally, an obvious best practice is to have the windowing operation applied once to the SQT matrix rather than to every input frame.

$$x_0 = 1/\cos\left(\frac{\pi}{f_s} \cdot \frac{f_0/2}{|2\kappa - \sigma|}\right)$$

$$A = 20 \cdot \log_{10} \cosh(2c \cdot \cosh^{-1}(x_0)) \quad (8)$$

$$W_g[u] = \exp\left(-\frac{(u-c)^2}{cf_s/f_0}\right) \quad (9)$$

3.4 Cepstral Measurement

To derive the harmonic sampling, first apply the Fourier transform to Equation 6 to obtain Equation 10, where $W_m[f, f_0]$ is the window's frequency response. (Let $W_m[f, f_0] = W_m[f]$ and $H[f, f_0] = H[f]$ since the non-linearity of H is discussed later in Section 5). After rearranging the terms, the glottal frequency response is Equation 11.

$$S[f] = 2\pi f W_m[f, f_0] * \underbrace{\sum_{m \in \mathbb{Z}} \delta(f - mf_0) e^{j\varphi} H[f, f_0]}_{G[f]} \quad (10)$$

$$G[f] = \sum_{m \in \mathbb{Z}} W_m[f - mf_0] = \frac{S[f] \cdot e^{-j\varphi}}{2\pi f \cdot H[f]} \quad (11)$$

The formants are much wider than the harmonic samples in G , and so the energy of H is approximately non-varying within $\pm \frac{f_0}{2}$. Therefore, $H[mf_0] \approx H[mf_0 \pm \frac{f_0}{2}]$, and at $f = mf_0 \pm \frac{f_0}{2}$, the $S[f]$ becomes the tract signal $H[f]$ plus some noise. The additive noise is sizable in the signal-to-noise (SNR) ratio because the impulse response is attenuated at that region. That is, $H[mf_0 \pm \frac{f_0}{2}] = S[mf_0 \pm \frac{f_0}{2}] \pm \epsilon$, and the spectral energy at mf_0 is

$$\begin{aligned} G[mf_0] &= \max_{\varphi} \frac{S[mf_0] \cdot e^{-j\varphi}}{2\pi f_0 \cdot H[mf_0]} \\ &\approx \max_{\varphi} \frac{S[mf_0] \cdot e^{-j\varphi} / 2\pi f_0}{S[f_0(m - \frac{1}{2})] \cdot e^{-j\varphi} + \epsilon} \end{aligned} \quad (12)$$

Since the detection of f_0 depends on the M harmonic observations, the detection probability $Pr[f_0]$ is equivalent to $\prod_m Pr[mf_0]$. Likewise, the cepstral similarly, $P[f_0]$, is equivalent to $\prod_m P[mf_0]$. The calculation is usually preferred in decibels (dB) for several reasons. Also the adjacent harmonics should be multiplied to filter out the undertones. The addition operation is computationally more safe than multiplication, and underflowing can be avoided with a stabilizing $\epsilon = 0.001$. The logarithmic option is also more time efficient than the multiplication when the log conversion of the matrix entries is applied in parallel. However, the log operation can be substituted by a fractional exponent. One of the possible cepstral measures $P[f_0]$ is therefore defined in Equation 13. Overall, the intensity of the quefrency correlates positively with $\prod_m |S[mf_0]|$ and negatively with $\prod_m |S[(m - 0.5)f_0]|$. The relation between them can be either a signal-noise subtraction or a SNR, and the subsequent detection can be done with either maxima or minima. This section covered the basic mathematical intuitions, and the next section uses the inferences.

$$\begin{aligned} P[f_0] &= \prod_m |G[mf_0]| \\ &\equiv \sum_m \pm \max_{\varphi} \{ 20 \log_{10} |\epsilon + S[mf_0] e^{-j\varphi}| \\ &\quad \mp 20 \log_{10} |\epsilon + S[f_0(m - 0.5)] e^{-j\varphi}| \\ &\quad \mp 20 \log_{10}(2\pi f_0) \} \text{ (dB)} \end{aligned} \quad (13)$$

4 Methodology

Having had the signal model of Equation 6, the quefrency measure in Equation 13, and the context of the previous section, this section applies the SQT matrix T , which is applicable on the frames $s[u]$, to obtain the desired speech features $S[f]$.

4.1 Notation

Let $(N + 1)$, M , and $(2c + 1)$ be the numbers of quefrencies, harmonics, and time samples respectively. The variables N , M , and c are user-defined integers larger than

one. Also f_{min} and f_{max} are variable f_0 boundaries, where $f_{min} < f_{max}$. Additionally, let $d \in \{1, 2, 4\}$ be a selectable phase synchrony mode, where the complexity levels 1, 2, and 4 correspond to real, complex, and sliding-phase types of detection. In terms of notation, let $\underline{s}_{(t,u)}$ be the u^{th} time sample in the t^{th} frame of N frames, and so \underline{s} is an $(N + 1) \times (2c + 1)$ matrix. In Equation 13, label the first cepstral signal S (of the modeled numerator) with $\kappa = 0$ and the second S (of the noise) $\kappa = 1$. Then, the first and second $S[f]e^{-j\varphi}$ of the equation can be expressed in a 5-d matrix, Equation 14. Now, an entry of S is equivalent to a vector product between a time frame and a slice of the transform. That is the matrix multiplication $S = \underline{s} \cdot T$. The next section describes how to build and use a transform such as the one in Figure 4.

$$|S_{(t,n,m,\omega,\kappa)}| = |\underline{s}_{(t,\forall u)} \times T_{(\forall u,n,m,\omega,\kappa)}| \quad (14)$$

4.2 Transformation

We define the T of SQT and its instantaneous frequency $f_{(n,m,\kappa)}$ as follows.

$$T_{(u,n,m,\omega,\kappa)} = \frac{W_{(u)} \cdot \cos\left(2\pi \frac{(u-c)}{f_s} f_{(n,m,\kappa)} + \varphi_{(\omega+d)}\right)}{\sum_u |W_{(u)}|} \quad (15)$$

$$f_{(n,m,\kappa)} = R_{(n)} \cdot (1 + m + \psi(R_{(n)}) - \kappa/2) \quad (16)$$

The R and φ are N -lengthed and d -lengthed vectors, obtained from the scale definition in Equation 3 and the instantaneous phase definition in Equation 7. W is a $(2c+1)$ -lengthed vector of a Chebyshev window with an attenuation of Equation 8. The frequency-shift function ψ is defined in Equation 21. The time index u was shifted by c and the phase index was shifted by d for filter centering. Table 1 completes the definition of the transform. In practice, the transform is constructed using matrix manipulations on Mesh Grid coordinates. It may also be necessary to flatten (or vectorize) the dimensions 2-5 in low-level programming languages.

For the general description of the transform, Figure 4 depicts the real part and imaginary parts of a flattened matrix example, $T_{(u,\dots,\kappa)}$. The depiction visualizes a stack of scallops rather than cochleas. The real part is even with respect to time, while the imaginary part is odd, so it is symmetric about the time origin. The vertical axis is the frame's interval in seconds. The first set of 48 filters along the horizontal axis corresponds to $\kappa = 0$ (the numerator S), and the remaining set of 48 filters corresponds to $\kappa = 1$ (the denominator). The difference between the two sets of functions is the scale with respect to time. In the two sets, the first 16 filters are of $m = 1$ (the first harmonic, the f_0), the next 16 filters are of $m = 2$ (the second harmonic), and the last 16 filters are of $m = 3$. The 16 filters in each are $n \in \{0, \dots, 15\}$. In the general case, the first axis corresponds to the time samples of windowed sinusoids. The second dimension varies the sinusoids' frequencies according to the quefrequencies, and the third varies per the harmonic overtones. The last two dimensions are for detection enhancement via phase synchronization and alias cancellation. The transform is applied on stationary

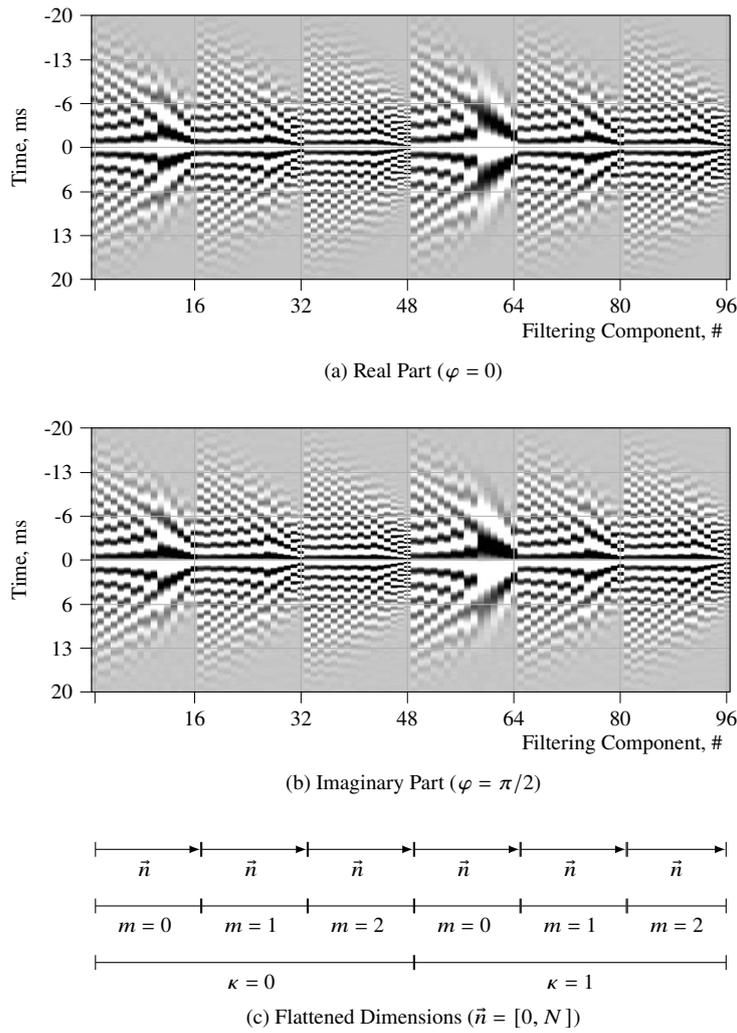


Fig. 4 SQT Matrix

frames. The output of the transform is a stochastic process that conveys three random variables: the periodicity, the envelope, and the noise. They are necessary to detect the quefrency f_0 , which shifts the modulating signal H_m .

4.3 Speech Feature Extraction

The sampled time frames are projected onto a three-dimensional feature space during the transformation. (See Equation 14). We theoretically demonstrated that including two extra axes (i.e., ω & κ) was a necessity for reliable quefrency readings. To obtain

```

Input:  $f_s = 8000, f_{min} = 100, f_{max} = 300, N = 15, M = 3,$ 
          $c = 160, d = 2, \sigma = 1.0$ 
Output: Transform Matrix, Scale Array
1 Initialize  $T$  as a  $(2c + 1) \times (N + 1) \times M \times d \times 2$  matrix
2 Initialize  $f$  as an  $(N + 1) \times M \times 2$  matrix
3 Initialize  $R$  and  $W$  as two  $(N + 1)$ -lengthed vectors
4 for  $n \in [0, N]$  do
5     Determine  $R_{(n)}$  (Equation 3).
6     for  $m \in [0, M - 1]$  do
7         for  $\kappa \in \{0, 1\}$  do
8             Determine  $f_{(n,m,\kappa)}$  (Equation 16).
9             if  $f_{(n,m,\kappa)} \leq f_s/2$  then
10                Determine  $W_{(n)}$  given frame-length of
                     $(2c + 1)$  and main-lobe width of
                     $0.5R_{(n)}/|2\kappa - \sigma|$ .
11                for  $u \in [0, 2c]$  do
12                    for  $\omega \in [0, d - 1]$  do
13                        Determine  $T_{(u,n,m,\omega,\kappa)}$ 
                            (Equations 15 and 21).
14 return  $T, R$ 

```

Table 1 The SQT Definition

the visioned 3-D speech space, apply absolute magnitude, synchronize by either a max operation or a summation along the ω axis, as in Equation 17, and extract the $\kappa = 0$ matrix slice. The $S_{(\bullet,n,m,\kappa=0)}$ resembles a monochromatic 2-D time-frame, where the n and m are the indices for the quefrequency rate track $f_0(t)$ and the normalized spectrograms $H_m(t)$.

$$S_{(t,n,m,\kappa)} = \max_{\omega} |S_{(t,n,m,\forall\omega,\kappa)}| \quad (17)$$

To calculate the cepstrogram measurements of Equation 13, apply a log or fractional exponent operator, flip the sign where $\kappa = 1$, and aggregate the κ and m terms. The log operator of the noise-canceling stage is not required, but pre-processing it with a 4-adjacency filter can bridge the transitioning between the connected pixels instead of significantly distorting the extracted information with the smoothing operator at the end.

$$P_{(t,n)} = \sum_{m,\kappa} (-1)^\kappa \cdot 20 \log_{10} S_{(t,n,\forall m,\forall \kappa)} \text{ (dB)} \quad (18)$$

The f_0 -levels (or indices) with the highest score estimate along the n -axis are extracted with the arguments of the maxima (*arg max*), as in Equation 19a. Given the f_0 track, the vocal tract responses (H_m) are extracted by applying the indices to Equation 19b. The f_0 -levels at t/f_r s are, then, converted to the Hertz unit by the R-Scale of Equation 19c, since the f_0 axis is quantized earlier per the reciprocal frequency scale R (of

Equation 3). Finally, the f_0 and H_m are the extracted speech features.

$$r[t] = \underset{n}{\operatorname{argmax}} [P_{(t, \forall n)}] \quad (19a)$$

$$H_m[t] = S_{(t, n=r[t], m, \kappa=0)} \quad (19b)$$

$$f_0[t] = R_{(n=r[t])} \quad (19c)$$

5 Analysis

and two-competing-methods formal Matlab implementations is given in

A preliminary comparison between three pitch extraction implementations is in Table 2. The Quefrency Transform Twelve (QT12) is one of our Matlab implementations based on the SQT method, which is defined in Table 1. The other two pitch extractors are formal Matlab implementations based on the Pitch Contours (PC) and the Amplitude Compression (AC), which are briefed in Section 2. According to Matlab, the two implementations are not entirely based on the proposals "A Pitch Estimation Filter Robust to High Levels of Noise (PEFAC)" and "Automatic Speaker Recognition Based on Pitch Contours" by Gonzalez et. al. and Atal respectively.

GPE (Gross Pitch Error) is a common speech metric of pitch performance. Smaller GPE value correlates with better methods. It is the probability that certain error rate is exceeded. GPE-20 is the probability of obtaining absolute error exceeding the significance level of 0.20 or 20% of the target label. The statistical significance validation was applied to the Matlab reference audio file "Counting-16-44p1-mono-15secs" with 0dB additive noise of "Turbine-16-44p1-mono-22secs." The SQT matrix was with 12 harmonic components; hence, QT12. Its complexity was adjusted so that it equals the sum of the complexities of the MFCC's and PC's implementations, since the SQT features are comparable to the features of both MFCC and PC. The table shows that the time complexity of the QT12 was lower than the AC's complexity and higher than the PC's. Also, the GPE of QT12 was lower than the PC's error and slightly higher than the AC's. Generally, the FFT spectrogram had the lowest computational complexity. A detailed evaluation is in Section 6. The features of FFT and the QT12 are recoverable. Just as FFT, the SQT technology can be utilized for AI speech composition and speech synthesis. Finally, the features of MFCC and FFT do not produce pitch tracks.

The remainder of this section provides technical elaborations in retrospect to the previous section. A case study highlights the findings, the end product, and the significance of the fundamental frequency for Natural Language Processing.

5.1 Speech Depth

The medium of the air particles is equivalent to a low-pass channel; it attenuates the high frequencies of the utterances. Speech audio is also anti-aliased (or low-pass filtered) while it is acquired and stored. Because higher FFs spread the cepstral code to higher frequencies, each speech depth naturally has a speech resolution, not to be confused with the frequency resolution. The depth must be one of the f_0 coordinates.

Features	Recoverability	Complexity, s	GPE-20, %
PC	N/A	0.046	6.589
AC	N/A	0.127	5.448
QT12	Yes	0.058	5.740
MFCC	No	0.012	N/A
FFT	Yes	0.007	N/A

Table 2 Validation Results of Preliminary Data

The speech code correlates with the shape of one period of the fundamental waveform. It can be modeled as $h(t, f_0)$ on the time domain, in which it exhibits a variation of a sine cardinal function, i.e., sinc, or modeled as $H(f, f_0)$ on the frequency domain, in which it can be approximated by Gaussian Mixture Models. The variability of the speech depth may have been fostering communications within households' members and facilitating language learning along one of the speech dimensions, since each speech depth happened to have a speech resolution and a cepstral band, and since a simple projection can align the speech features of different speakers' characteristics. The speech depths are commonly associated with masculinity, femininity, and infancy. However, the overall majority of human beings naturally produce the resolution that is also a common characteristic of the youth voice.

Newborns first encounter blurred speech, and the low resolution perhaps helps humans acquire and model languages in a gradual general-to-specific heuristic search, guiding the internal neural systems. Generally, increasing f_0 prioritizes the voice over the voices with lesser ones; this is ascribable to the voice masking phenomena, which is caused by the fact that two time periods of a high f_0 resemble a time period of a low f_0 , but not vice versa. (The utterances consist of pulse wavelets rather than sinusoids). For example, requiring more signal processing filters, listening to a conversation whose background is voices of unsatisfied dependents requires extra mental work to filter out high f_0 interference. The unpleasant noise may further the survival of the species, preventing child neglect. On the other hand, lowering the f_0 expands its voice coverage since signals traveling on lower frequencies optimize their energy for long distance communications (like in open fields). Although the gender may have a degree of correlation because of the vocal folds' lengths Fitch (1997), the association does not hold because of the existence of sizable minorities, which make the majority of human voices have high pitch. Moreover, the voice generator can transit between the speech resolutions regardless of sex and age. For instance, parents tend to use infant-directed speech to promote communication and learning. Recent findings showed that infants who babble at early age receive contingent feedback from social interactions that foster language learning Albert et al. (2018). Likewise, infants can produce adolescent-directed speech. The illustration of Figure 1 about the phenomena affirms previous speculations in the literature. For example, Hoeschele Hoeschele (2017) and Warren et al. Warren et al. (2003) mentioned the consensus of the two-dimension view of f_0 was motivated by auditory expertise.

One may speculate about the underlying physical constraints that prompt the λ_0 doubling; however, the teleport path, which is shown in Figure 3, indicates that the f_0 has an additional dimension. A rotation around the second axis is called the height of the pitch and is believed to be $height = 110\text{Hz}$ although this number is not exactly the same for every human being, especially with the environmental influence. A 2017 survey by Bernhardsson Bernhardsson (2017) in Github showed that the arithmetic population mean of f_0 varies per language. Moreover, similar to speech production, human perception of speech does not considerably discriminate between the intonations that are one octave apart. The term pitch implies the angular position, while the f_0 refers to the measurement reading. The relation between the two terms is expressed mathematical in Equation 20: the f_0 is congruent to the *pitch* modulo *height*.

$$\begin{aligned} f_0 &\equiv pitch \pmod{height} \\ &= pitch + i \cdot depth \Big|_{i = height} \\ &\quad depth \in \{0.5, 1, 2, 4\} \end{aligned} \quad (20)$$

5.2 Case Study

Figure 5a prints the spectrogram of four Wake-Up-Word (WUW) utterances from the WUWII Corpus Kepuska (2011). The utterances are, from left to right, a female voice of "Voyager," a male voice of "Voyager," another male voice of "Operator," and another female voice of "Operator." The last utterance ends with a noticeable tone. Each waveform of the four was normalized by its extrema before they were concatenated. Comparing the H_m (i.e., timbres) of the four utterances, one may notice a shift in the intensities of the overtones. Based on the shift within the Mel-scale, a harmonic shift ψ can be approximated as in Equation 21. Figure 5b portrays a partition of the frame data. Applying the transform (T) on the frames s (as in Equation 18) gives the normalizing sampled spectrogram S and the cepstrogram P . Applying Equations 19a and 19c on the cepstral similarity extracts the f_0 path.

$$\psi(f_0) = \max\left\{0, \left\lfloor \frac{700}{f_0} - \pi \right\rfloor\right\}, \quad (f_0 \text{Hz}/s') \quad (21)$$

Figure 6 renders two cepstrograms of two configurations: fast (Figure 6a) and boosted or refined (Figure 6b). The common parameters of the two cases are $f_{min}:100$, $f_{max}:300$, $c:160$, $d:2$. The first option is economical in terms of its resource consumption (and faster than the latter). The configuration parameters of the first are $N:7$, $M:3$, $\text{Shift}:0$, $\sigma:1.5$. Observing three overtones ($M = 3$) turned out to be adequate for the pitch estimations (in a low-noise setting). The size of its matrix transform is 321×96 . On the other hand, the parameters of the boosted configuration, whose transform size is 321×320 , are $N:15$, $M:5$, $\text{Shift}:\psi$, $\sigma:1.0$. Considering more harmonic observations ($M = 5$), the second generates f_0 readings that are sharper than the readings of the fast configuration.

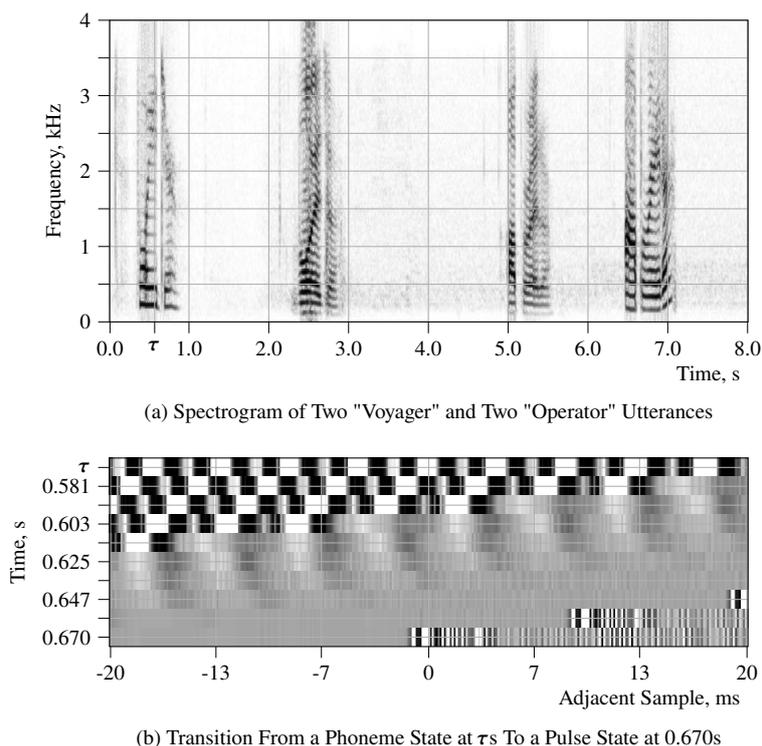
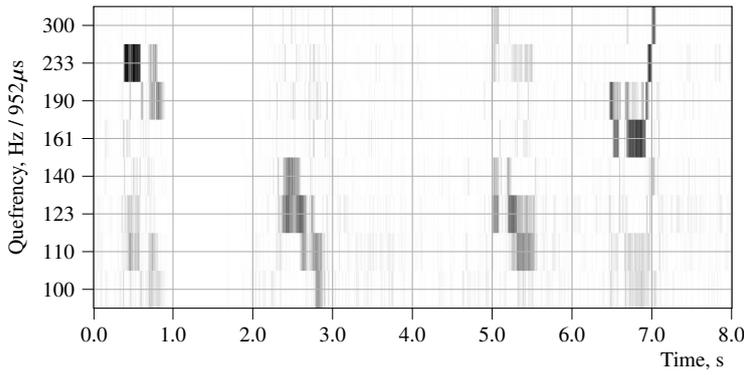


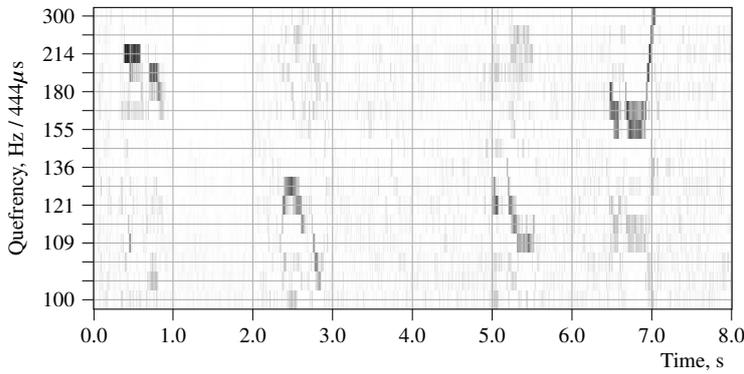
Fig. 5 Illustrative Example from the WUW Corpus

The detection of the pitch track in Figure 6a has larger variances than the pitch track in Figure 6b. The large variances are due to the reduced quefrequency resolution. Reducing the number of bins ($N + 1$) reduces the time complexity. Meanwhile, the refined results in the boosted case have a relatively large number of quantization levels, increasing the precision of the readings. It is worth noting that the refined cepstragram can be composed of two smaller ones; hence, the output quality can scale up with dynamic programming or boosting techniques. Additionally, one may notice that speech closure pulses can be located from the sharp pitch patterns, and so averaging (smoothing) the pitch track was avoided so as not to distort the speech signals. Moreover, the path can be rendered discontinuous in the unvoiced intervals by applying a voice activity detection that utilizes the extracted pitch energies along the vividest track.

The extracted cepstragrams illustrate that the voiced samples of different depths are linearly separable in the quefrequency domain. Two speakers utilized the quefrequency channel between 155 and 300Hz/s', and two other speakers utilized the channel between 100 and 155Hz/s'. In another similar example, the first and last two seconds of the case study (Figure 5a) were added to the babbling (Figure 1a), and a high definition quefrequency transform (sized 321×4096 with $N:63$, $M:16$, and $\sigma:0.5$) filtered



(a) Economical Case



(b) Boosted Case

Fig. 6 Cepstrograms' Complexities

the utterances from the congested audio signal that has three voices at a time. The cepstragram of that example is demonstrated in Figure 2b. The intelligibility of the recovered speech correlated with the f_0 as was expected.

Nevertheless, based on Parseval's theorem, the total energy in the time domain is comparable with the aggregated spectral energy, as in Equation 22. The Root Mean Square (RMS), preferred for describing the audio time-frames, is the square root of energy. The theorem holds true only when the applied windows are rectangular, so the proximity in the equation occurs when the frames and the transform are windowed. Also, the right hand side is the L^2 -norm of H_m since the denominator is already considered in the windowed transform (Equation 15). Figure 7 plots the two sides of the proximity to illustrate the energies of the extracted pitch tracks. It shows that the increased number of harmonics along with the harmonic shift ψ increased the extracted pitch RMS and energies. The voice activity can be detected from the boosted case (Figure 7b) better than from the swift extraction (Figure 7a).

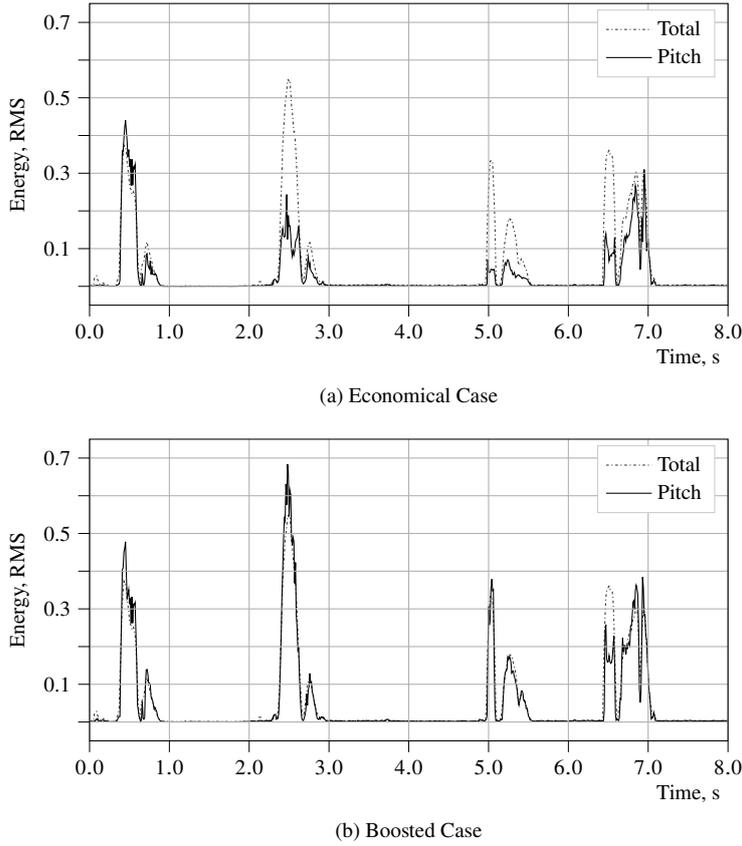
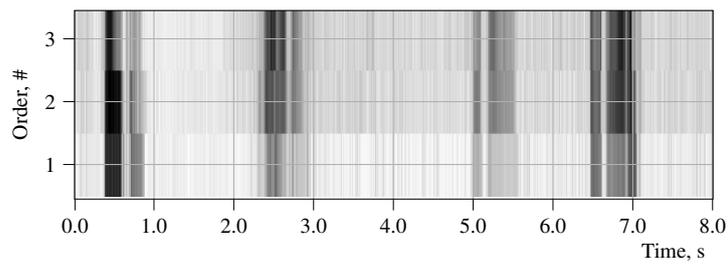


Fig. 7 Energies of Total and Pitch Extractions

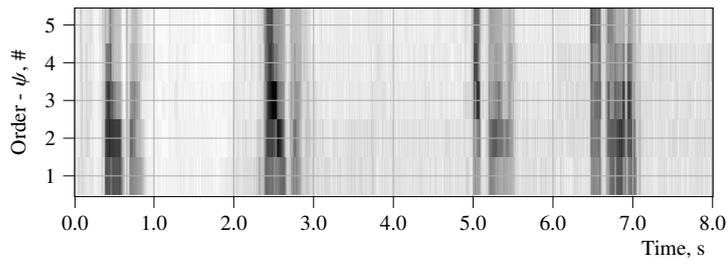
The extracted samples, which compose a harmonically normalized spectrogram H_m , is depicted in Figure 8 for each of the two cases. Given the signal model (of Equation 6), $f_0[t]$, and the two $H_m[t]$, speech signals were reconstructed, and their regular spectrograms are depicted in Figures 9. The depictions show that the extracted harmonic elements were placed back to their spectral locations. One can see that the discrete shift skips a few harmonic elements in the second and third voices to normalize the frequency scaling of the vocal tract. It is worth noting that the correlation between the f_0 and the tract is a correlation between two glottal characteristics.

$$RMS_{(t)} = \left| \sum_u s_{(t, \nu u)}^2 / (2c + 1) \right|^{\frac{1}{2}} \approx \left| \sum_m H_{(t, \nu m)}^2 \right|^{\frac{1}{2}} \quad (22)$$

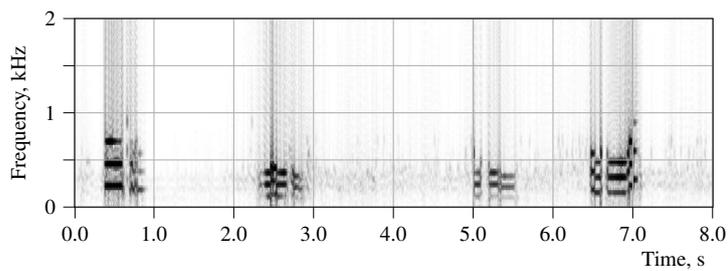
The more harmonic overtones, the higher the speech quality. If the minimal frequency of the speech model is $f_{min} = 50\text{Hz}$, the maximal number of overtones within the bandwidth $\frac{f_s}{2} = 4\text{kHz}$ is $\lfloor \frac{f_s/2}{50} \rfloor = 80$. Figure 10a depicts all H_m components in



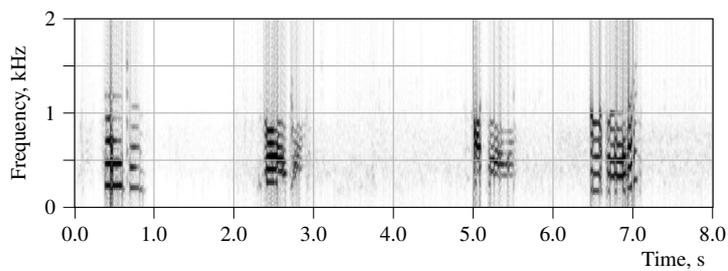
(a) Economical Case



(b) Boosted Case

Fig. 8 Spectra Features (H_m)

(a) Economical Case



(b) Boosted Case

Fig. 9 Spectrograms of Reconstructed Speech Signals

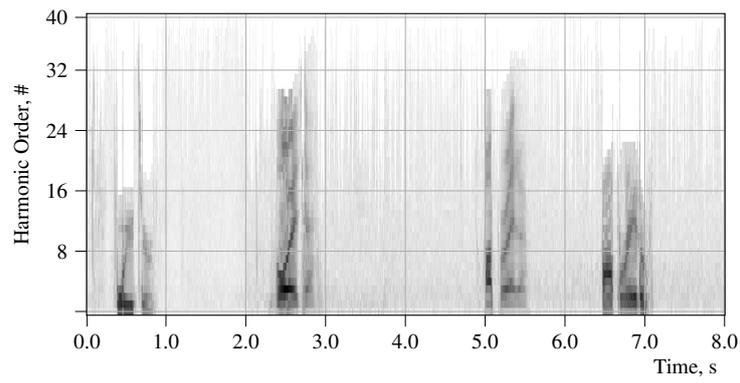
4kHz, and its corresponding spectrogram reconstruction is in Figure 11a. It recovers a spectrogram approximation of the original input signal (Figure 5a). Another configuration is to subsample the speech spectral code to $\lfloor \frac{f_s/2}{330} \rfloor = 12$ features as in Figure 10b. This reconstruction can be intelligible with a large number of bins.

The first two speech formants are within the first 16 overtones, as shown in Figure 11a. Needless to say, the significance of the harmonic digit decreases as the harmonic rank increases. Also, the f_0 is included in the figure as the smallest tone. Since the first two utterances correspond to the wake-up-word "Voyager," while the last two do so to "Operator," one can conclude that the high depth adds extra speech details that may not be essential in differentiating between the two speech utterances. Consequently, only the first speech features are usually considered. For example, the MFCC method includes the first 13 cepstral coefficients. However, in the standard MFCC procedure, the overtones with high ranks are averaged or interfered. This is because its Mel-Scale is applied to the downsampled spectrograms, shown in Figure 11b. For that reason, the MFCC is not aligning the harmonic ranks. In the figure, the eight harmonics of the first utterance are mapped to the fifteenth harmonics of the second utterance. One possible normalization is depicted in Figures 10c and 11c. The normalization aligns the similar features of the similar utterances by considering an equal number of harmonics. Additionally, the ψ function increased the normalization even more, and it became more resilient to the quantization error than the subsampling of Figures 10b and 11b. Finally, digitized 8-level H_m and 32-level f_0 may be sufficient for machine learning. If the number of harmonics (M) is 12 and the frame rate is $f_r = 100\text{fps}$, the two levels aggregate to $(M \cdot \log_2(8) + \log_2(32)) \cdot f_r = 4.1\text{kbps}$ (kilobit-per-second), which is a reduced transmission bandwidth.

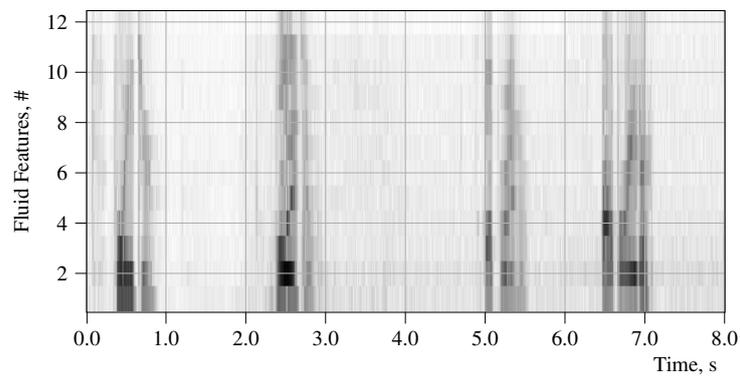
5.3 Spans of Language and Intelligence

The generation and recognition of spoken languages are sophisticated processes. Since the intelligence of a species can function as a means of its survival, the two processes can be genetically optimized during a lengthy selective reproduction phase, as languages interconnect with and boost intelligence. For example, having been bestowed control over their own breaths, species, like dolphins, elephants, and birds, have been able to extend their senses beyond their lines of sight and share alerts and information within the kind. For instance, whales communicate at long distances and navigate their surroundings, transmitting sound units and receiving sonar echos. The utilization of the spoken units is a founding module for intelligence just as the human utilization of written alphabetic and numerical symbols is a basis for written knowledge, commerce, and civilization. Human languages come in several forms, and each has a countable number of units (or letters). In the spoken one, the phoneme is the unit of speech.

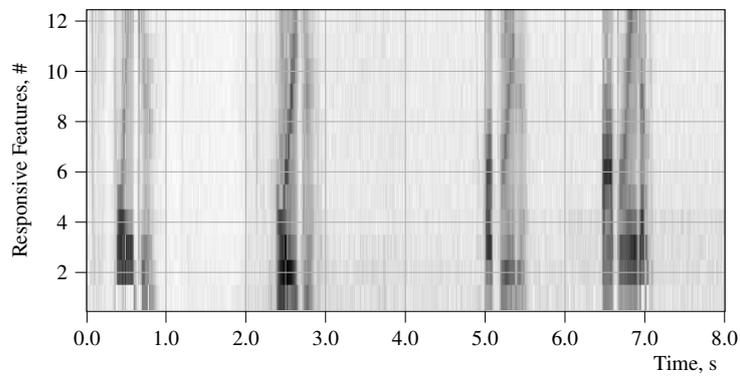
The ability of comprehending logic and sequential series of events must have been, to a certain degree, built upon the primal ability of recognizing sensory data, one of which is speech. The former is equivalent to the latter when the domain of the speech spans multiple days as opposed to minutes. Some may argue that some individuals learn to walk first while some talk first. Even so, human language acquisition begins

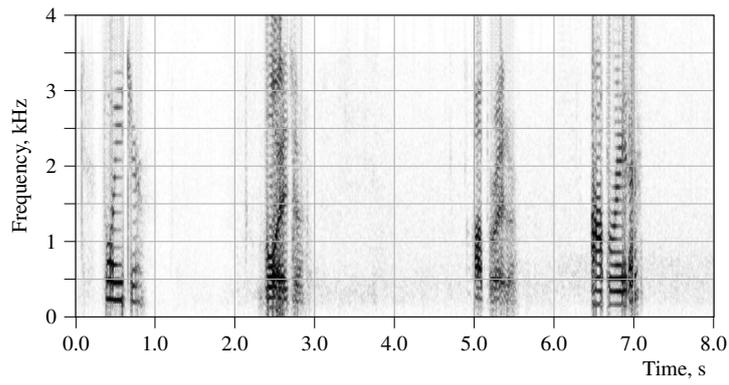


(a) In-Bandwidth Harmonics

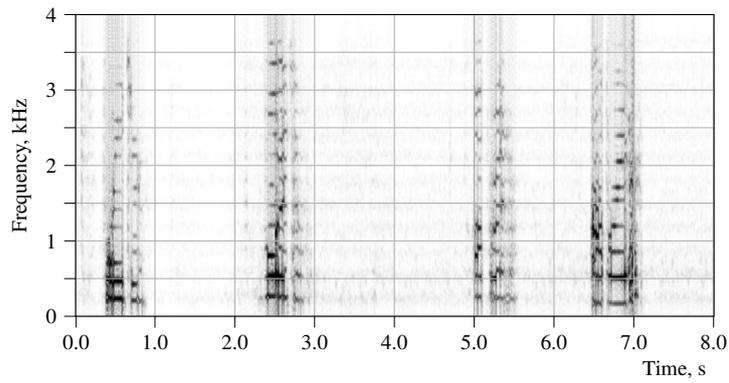


(b) Enhanced Spectrogram

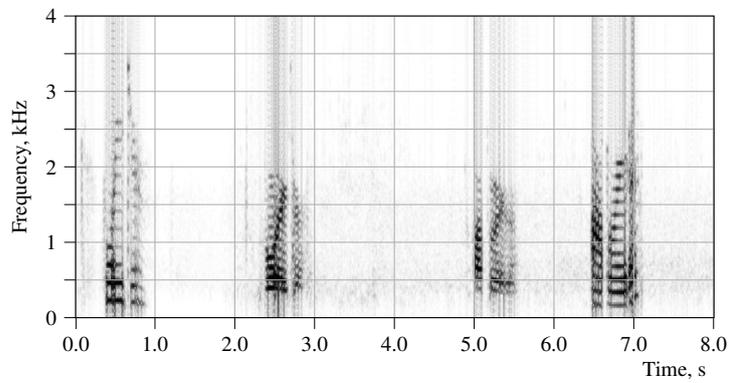
(c) Normalized Spectrogram (H_m)**Fig. 10** Comparison Between Frequency Sampling Configurations



(a) In-Bandwidth Harmonics Case



(b) Enhanced Spectrogram Case



(c) Normalized Spectrogram Case

Fig. 11 Spectrograms of Reconstructed Speech Corresponding Respectively to the Cases in Figure 10

much earlier with quasi-resonant vocalizations, according to Psychologist Rachel Albert (2013). Moreover, according to Linguist Noam Chomsky (1965), there exists an innate Language Acquisition Device (LAD) that prepositions concepts, such as nouns and verbs, in the human languages. It is true that the tuneful pattern is not exclusive to humans; however, the speech ability has been vital to humankind Pisanski et al. (2016). Our hypothetical reasoning is that graceful recognition of temporal sequences also enabled the cognition of long-term chronological occurrences, which, when optimized (in succeeding iterative genetic mutations), intellect possibly emerged naturally in the species. Constituting a self-aware agent, hierarchical spans of language are levels for intelligence. Increasing the contrast of visualization can extend the attention spans hindered due to developmental disorders Asiry et al. (2018). Additionally, magnetic resonance imaging (MRI) has revealed that both the processing of language and the ability of using tools stimulate similar neural areas Higuchi et al. (2009). Undoubtedly, speech signal processing is part of decision-making processes.

Having been blessed with automation, human populations flourished. For instance, individuals merited labor rights, when steam engines were employed in the agriculture industry, but the recent technological leaps may require safe domestic machines that add the human component to the autonomous world. Since the acoustic agents (or companions) are expected to comprehend meanings and interact not only intellectually but also to form bonds with human agents, the artificial agent would first have to receive and generate a voice in a similar manner to humankind, although not necessarily in biological means. Artificial systems equipped with digital front ends can have a human-like learning phase as well as subjectivity and perhaps artificial feelings.

5.4 Paralinguistic Intonation

Speech utterances sound natural when they are carried on the f_0 pattern, which can be regarded either a discontinuous or a continuous discrete function $f_0[t]$, as in Figure 12. The fundamental frequency (f_0) is the main speech feature, vital in natural language processing, especially in languages where speech stresses play a major role in defining the speech parts and grammar. Additionally, modeling English grammar would lean on the $f_0[t]$ patterns.

Intonation patterns, sometimes marked with diacritics, are essential in word recognition and are common in Eastern languages. According to Albert Mehrabian Mehrabian (1972), the nonverbal human tone carries shades of meanings. For instance, the only difference between the two Korean words: "악다" (sit) and "앉다" (hug) is that the emphasis makes the second "anta" slightly longer. Other examples are the tashdīd emphasis in Arabic, the acute accentuation in Greek, and the compound words in English. For instance, the stressed syllable in "thermometer" is just as important as its phoneme sequence. Moreover, the regular pattern of the f_0 is a reduction since it correlates with the breathing pattern. For instance, adults breath slower and so naturally do their glottises. The general pattern was shown in previous publications, according to which also the emotions, such as happiness and sadness, correlate with the pitch pattern Green et al. (2011). Utterances normally de-accelerate toward local minima. In contrast, f_0 increments appear while appending upcoming expressions.

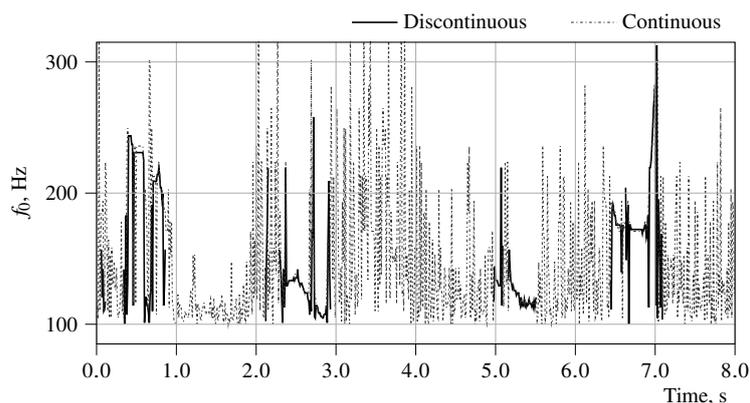


Fig. 12 The Pitch Tracks of the Samplers in Figures 10a (Discontinuous) and 10c (Continuous)

The f_0 function holds clues for several grammar components: punctuation periods, clauses, and stresses. A persistent upward trend may imply a preparation; the accelerating is salient in the regularly de-accelerating pattern. Exclamation points and question marks are slightly similar; the former is a (linear) trend across the utterance, while the latter is an f_0 suffix (usually an exponential one). Additionally, the speech emphasis is a short up-down bounce, which is the most frequent pattern. A varying f_0 can possibly grasp attention as the auditory focuses of the listeners are more likely to intersect with the speaker's tone. Such an intersection maximizes the reception of the speech. That is, words uttered with bouncing tones do sound emphasized as they widen in the quefrequency domain. Ample pitch patterns would have to be analyzed for conclusions.

The syllables of verbs are emphasized differently than those of nouns. The utterances of neutral statements and wake-up-word requests diverge naturally in the f_0 pattern. The f_0 highlights some parts of speech, and this is crucial for machine learning and language understanding. This is because the f_0 is a primary component in natural languages.

6 Pitch Extraction Evaluation

6.1 Results

Table 3 prints out performance metrics of three methods under several noise conditions, totaling 28 tests. QT mostly had the lowest probability of Type-1 error when the significance thresholds were set to 20% and 10%. QT and PC were neck and neck when the threshold was set to 5%. The proposal evidently outperformed the other two methods in all of the additive noise conditions that were tested: noise-free, white-noise, and turbine-noise. AC generally performed well when the turbine-noise level was high, and PC generally did so when the white-noise level was low. The

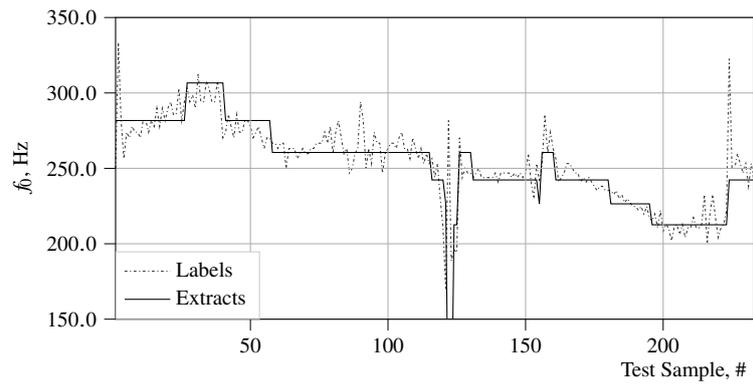
Settings		Methods	Error Metrics						
			Lag	GPE-20	GPE-10	GPE-05	MSE		
No Noise		QT	0.00	2.18	5.84	14.34	647.3		
		AC	0.0	4.34	8.03	18.28	2101.7		
		DC	0.0	3.65	7.88	15.77	1205.9		
White-Noise		20dB		QT	4.97	2.24	5.90	14.42	663.9
				AC	11.0	4.40	8.10	18.29	2104.7
				DC	0.5	3.66	7.91	15.82	1193.7
		10dB		QT	0.97	2.66	6.30	14.82	743.23
				AC	37.65	5.13	8.92	19.08	2389.5
				DC	29.87	3.91	8.28	16.38	1185.4
		0dB		QT	3.00	6.74	10.32	18.81	1496.2
				AC	56.4	11.89	15.63	26.30	5617.0
				DC	66.3	9.15	13.67	23.25	2022.4
Turbine-Noise		20dB		QT	4.39	2.46	6.13	14.61	646.19
				AC	12.5	4.66	8.39	18.51	2094.7
				DC	16.1	3.80	8.09	16.00	1171.2
		10dB		QT	11.74	5.70	9.3213	17.532	1031.2
				AC	33.71	8.45	12.36	22.46	2625.4
				DC	59.71	7.55	11.89	20.23	1692.8
		0dB		QT	27.81	30.23	34.23	41.08	3783.2
				AC	35.3	30.34	34.99	45.68	6079.8
				DC	45.1	37.33	41.86	48.83	6451.8

Table 3 Test Results of the FDA Data

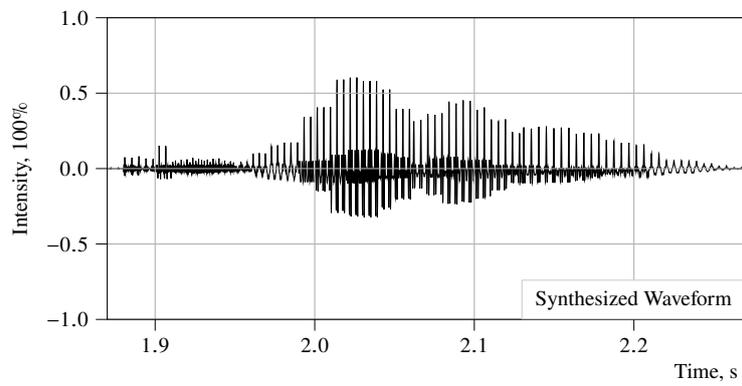
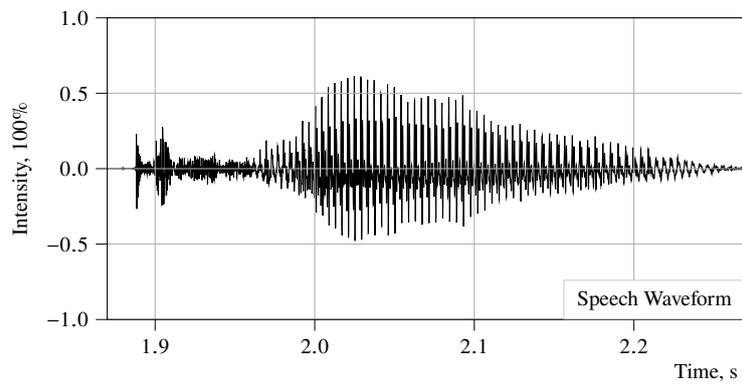
evaluation shows that the pitch detection performance of the proposed method performed excellently. Another indication for the pitch accuracy is the speech quality of the reconstruction. It may be safe to assume that QT is relatively robust in high noise conditions.

6.2 Test Samples

Figure 13a demonstrates the SQT approach in practice, whereby higher frequencies were assigned fewer quantization levels. Figure 13b plots the samples of two waveform signals: the original and its twelve-tone reconstruction. The composition is no small feat, and its RMS envelope is fair. The speaker traits, like accent, however, were less present, and that was expected since only a dozen components were extracted for this set of tests.



(a) Pitch Track



(b) Speech Reconstructicon of a "Car" Utterance

Fig. 13 Illustrative Example from the FDA Data

6.3 Experimental Settings

The selected evaluation dataset was "Fundamental Frequency Determination Algorithm (FDA) Evaluation Database" by Paul Bagshaw, Centre for Speech Technology Research, University of Edinburgh, UK. The FDA data aggregates 5.527 minutes of audio, sampled at $20kHz$, and labeled fundamental frequencies of male and female speakers. The audio was downsampled to $8kHz$ before the extractions' evaluations, and best lag alignment within double the window size was performed for each method and each test. The outputs of the methods were also post-processed using a median filter whose kernel size is three. The utilized platform was Matlab Online 2019b (9.7.0). The average feature-extraction times for the SQT, ACGonzalez and Brookes (2011), and PCAtal (1972) methods were 1.34, 3.24, and 1.16 seconds, respectively. The QT was constructed with Gaussian windows and $M = 12$, the first five of which were used for the pitch detection.

6.4 Error Metrics

The metrics are variant significance testing cases and a squared loss case; the gross pitch error (GPE), its lag, its significance (GPE-Threshold), and the mean-squared-error (MSE) are defined in Equation 23. Two types of noise signals, white noise and turbine noise, were added in three SNR cases: low (20dB), medium (10dB), and high (0dB) noise conditions. The metrics of the tests involving randomness are represented by the average of 30 repetitions; each were determined at the most suitable GPE lags.

$$\begin{aligned}
 \text{GPE} &= |\hat{f}_0 - f_0| \\
 \text{Lag} &= \underset{\Delta t}{\text{argmax}} |\hat{f}_0[t - \Delta t] - f_0[t]| / F_s, s \\
 \text{GPE-Threshold} &= P(\hat{f}_0 > f_0 \cdot \text{Threshold}/100), \% \\
 \text{MSE} &= |\hat{f}_0 - f_0|^2
 \end{aligned} \tag{23}$$

7 Conclusion

This work refines the quefrency definition as a unit of acceleration, measured in $Hz/\mu s$, and the f_0 is a primary speech feature, not separable from the spectral energies. While the commonplace frequency banking does not make the MFCC features reconstructable, the proposed method adjusts the frequency banking implicitly in the windows' the frequency responses to make the features reconstructable. The cepstral method is adequate when it is configured as in the proposed approach since the windowing operator convolves with or widens the frequency domain and attenuates its magnitudes exponentially. Furthermore, the proposed method calculates the pitch and its spectrum at the same time; it is more efficient than an equivalent combination of other state-of-the-art methods.

The findings confirm experts' speculations, such that the f_0 has two dimensions: depth and intonation. The two dimensions are considerable because speakers sometimes gear between the depths almost instantly. While the depth determines the speech

resolution, the intonation is crucial for natural language processing and Wake-Up-Word (WUW) systems since it communicates urgency and breath patterns. The speech resolutions are the result of the stationarity, the speech depths, and band-limited speech channels, limiting the number of the speech components.

The f_0 intonation clues primitive language. It may be evident that the language expressed in the intonation precedes the shaping capability of the vocal tract. Like facial expressions, the pitch patterns could be prehistoric and universal. The proposed approach addresses several challenges in order to elevate speech processing to the comprehending level. It is crucial that ASR systems effortlessly detect and compose speech in resolutions that conserve energy, optimize the features' SNR, and preserve the speech components during extraction.

The proposed Speech Quefrency Transform (SQT) is suitable for artificial intelligence processing. The proposed method achieved a relatively very low Mean Square Error (MSE) via the frequency demodulation assumption. The transform expands the speech samples into a hyperspace whose axes correspond to pitch, harmonic spectra, and frequency-based anti-aliasing. The dimension expansion was designed while the unsupervised machine learning and clustering are in mind. Whenever the ASR is attuned to an f_0 carrier, the quality of the extract is high since the background noise is attenuated. Distant- and multi-speech detection and extraction are expected to be more feasible with the provided method. For example, a heuristic acoustic model may cache recent utterances to gradually increase observations as language and contextual models require for an observation instance. For fast processors, the first five harmonica, in which the first formant usually resides, are sufficient. For speech normalization, the first twelve harmonica, conveying the first two speech formants, are so. The method is compatible with high performance computing since it consists of matrix operations. The SQT method attenuates the background noise, and its input can be reconstructed.

Acknowledgements Deep gratitude to Prof Judith B Strother for her proofreading review of the work.

References

- Albert RR (2013) Bidirectional influences of social feedback on parent-infant communication. PhD thesis, Cornell University
- Albert RR, Schwade JA, Goldstein MH (2018) The social functions of babbling: acoustic and contextual characteristics that facilitate maternal responsiveness. *Developmental science* 21(5):e12641
- Asiry O, Shen H, Wyeld T, Balkhy S (2018) Extending attention span for children adhd using an attentive visual interface. In: 2018 22nd International Conference Information Visualisation (IV), IEEE, pp 188–193
- Atal BS (1972) Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America* 52(6B):1687–1697
- Atlas L, Janssen C (2005) Coherent modulation spectral filtering for single-channel music source separation. In: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., IEEE, vol 4, pp iv–461
- Bernhardsson E (2017) lang-pitch. <https://github.com/erikbern/lang-pitch>

- Bogert B, Healy M, Tukey J (1962) The quefrency analysis of time series for echos. In: Proc. Symposium on Time Series Analysis, pp 209–43
- Chen N, Hu Y (2007) Pitch detection algorithm based on teager energy operator and spatial correlation function. In: 2007 International Conference on Machine Learning and Cybernetics, IEEE, vol 5, pp 2456–2460
- Chonsky N (1965) Aspects of the theory of syntax
- Clark P, Atlas L (2009) A sum-of-products model for effective coherent modulation filtering. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp 4485–4488
- De La Cuadra P, Master AS, Sapp C (2001) Efficient pitch detection techniques for interactive music. In: ICMC
- Ewender T, Hoffmann S, Pfister B (2009) Nearly perfect detection of continuous f_0 contour and frame classification for tts synthesis. In: Tenth Annual Conference of the International Speech Communication Association
- 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
- Gonzalez S, Brookes M (2011) A pitch estimation filter robust to high levels of noise (prefac). In: 2011 19th European Signal Processing Conference, IEEE, pp 451–455
- 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. *Emotion* 11(5):1124
- Higuchi S, Chaminade T, Imamizu H, Kawato M (2009) Shared neural correlates for language and tool use in broca’s area. *Neuroreport* 20(15):1376–1381
- Hoeschele M (2017) Animal pitch perception: Melodies and harmonies. *Comparative cognition & behavior reviews* 12:5
- Jurafsky D, Martin JH (2009) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Second Edition). Pearson/Prentice Hall Upper Saddle River
- Kawahara H, Masuda-Katsuse I, De Cheveigne A (1999) Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds. *Speech communication* 27(3-4):187–207
- Kepuska V (2011) *Wake-Up-Word Speech Recognition*, Speech Technologies. InTech Open
- Lathi BP, Green RA (2005) *Linear systems and signals*, vol 2. Oxford University Press New York
- Li Q, Atlas L (2008) Coherent modulation filtering for speech. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp 4481–4484
- Mehrabian A (1972) *Nonverbal communication*. Transaction Publishers
- 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
- Oppenheim AV, Schaffer RW (2004) From frequency to quefrency: A history of the cepstrum. *IEEE signal processing Magazine* 21(5):95–106

- Pisanski K, Cartei V, McGettigan C, Raine J, Reby D (2016) Voice modulation: a window into the origins of human vocal control? *Trends in cognitive sciences* 20(4):304–318
- Rabiner L, Schafer R (2011) Digital speech processing. *The Froehlich/Kent Encyclopedia of Telecommunications* 6:237–258
- 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
- Rabiner LR, Gold B (1975) *Theory and application of digital signal processing*. Englewood Cliffs, NJ, Prentice-Hall, Inc, 1975 777 p
- Smith JO (2011 edition) *Spectral Audio Signal Processing*. CCRMA Stanford
- Stefanatos GA, Green GG, Ratcliff GG (1989) Neurophysiological evidence of auditory channel anomalies in developmental dysphasia. *Archives of Neurology* 46(8):871–875
- Talkin D (1995) A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis* 495:518
- 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
- Wölfel M, McDonough JW (2009) Distant speech recognition. *Wiley Online Library*
- Zahorian SA, Hu H (2008) A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America* 123(6):4559–4571