CHAPTER 14

A Robust Algorithm for Pitch Tracking (RAPT)

David Talkin

Entropic Research Laboratory
Suite 202
600 Pennsylvania Ave. S.E.
Washington, D.C. 20003, USA

Contents

Speech Coding and Synthesis
Edited by W.B. Kleijn and K.K. Paliwal

1. Introduction

The need for reliable automatic estimates of the voice fundamental frequency has engaged as many creative minds as any topic in speech analysis. Reports on the algorithms and systems resulting from these efforts comprise a rich literature spanning several generations. The reader is referred to [1, 2] for a comprehensive historical summary and bibliography. It is not possible in a single book chapter to review all of the significant work in this area. Here, we will provide a sketch of human speech production mechanisms; define the problem; provide a very abbreviated survey of pitch-tracking techniques; and then then focus on a complete description of a robust algorithm for pitch tracking, RAPT, that has proven effective in the context of basic research and synthesis engineering.

1.1. What is pitch?

Strictly speaking, the term *pitch* should be reserved for the auditory percept of tone. This can be measured, for instance, by asking a listener to compare alternate presentations of a complex (multi-component) signal, the pitch of which is to be estimated, with a pure sinusoid of variable frequency. When the listener has adjusted the sinusoid so that it seems to be the same tone as the complex stimulus, the sinusoid frequency could be defined to be the *pitch* of the complex signal. Although some computational auditory models are successful at predicting the perceived pitch of certain complex signals, pitch is not directly measurable from the signal and is a nonlinear function of the signal's spectral and temporal energy distribution.

Fundamental frequency (F0) is the quantity that is being estimated by virtually all "pitch trackers". F0 is an inherent property of periodic signals, and tends to correlate well with perceived pitch. For the purposes of this chapter it is defined as the inverse of the smallest *true* period in the interval being analyzed. This definition provides for the short-time variation in F0 that is observable in human speech. As will be seen later, determination of "true" is the crux of the matter!

1.2. A glimpse at speech production

In order to understand the nature of the F0 estimation problem it is helpful to review quickly some of the human speech producing physiology and state some simplifying assumptions. See [3] and [4] for thorough coverage. The *glottis* is the opening in the larynx that is adjustable by the *vocal folds* ("vocal cords" in the vernacular). Muscles and cartilage in the larynx provide several dimensions of adjustment of the folds, including the degree of their V-shaped opening, their longitudinal tension, and the stiffness of their bulk. These adjustments are under more-or-less voluntary control. When the folds are adjusted appropriately and when airflow from the lungs via the trachea is of sufficient velocity, the folds self-oscillate. Note that although, to first order, the tensions and geometric configuration established by

the laryngeal muscular adjustments do not change in the course of a single cycle of fold vibration, adjustments of pulmonary effort and of laryngeal configuration are the primary determiners of the amplitude and rate of vocal-fold vibration on time scales corresponding to phone and word production.

As the folds oscillate, they vary the degree of glottal opening, which in turn modulates the volume of air passing through the glottis. It is this periodic airflow modulation that serves as the excitation for the vocal tract during *voiced speech.* Laryngeal and pulmonary adjustments also permit some voluntary control of voice *quality.* For instance, if the folds are pressed more forcefully together, they are more thoroughly closed during a longer fraction of the cycle and tend to open more abruptly. This tends to reduce the amplitude of the first harmonic and boost the amplitudes of higher harmonics in the resulting *pressed voice* speech signal. An increase in pulmonary pressure tends to cause higher amplitude oscillations and thus, more abrupt stoppage of the airflow as the folds slam together at the beginning of the *closed glottis interval* of the cycle. This, in turn, leads to a speech signal with overall higher amplitude and also differentially increases the spectral energy at higher frequencies. A decrease in pulmonary effort and/or a widening of the glottal opening leads to *breathy voice* associated with incomplete or soft glottal closure and manifested as speech having a relatively large aperiodic component and most of the spectral energy concentrated in the first few harmonics [5, 6].

The airflow through the glottis or the *glottal volume velocity,* $U(t)$, is the forcing function that ultimately determines the periodicity of voiced speech. Stylized versions of $U(t)$ and its derivative $U'(t)$ are shown in fig. 1. Also shown are estimates of the corresponding functions obtained by auto-regressive (linear prediction) inverse filtering of natural speech. The shape of the "normal" $U(t)$ waveform is explained as follows: When the vocal folds close, they close completely, and U is zero. The folds open due to the combined effect of sub-glottal air pressure and the stored energy in the fold's mechanical system. U rises very gradually because the folds open gradually and, perhaps more importantly, because of the inductance of the tracheal and superglottal system which prevents instantaneous acceleration of the air column. At some point, the mechanical restoring force in the folds causes them to begin to close. However, the inductance now tends to keep the air flowing. As the folds approximate, the Bernoulli force becomes significant, further accelerating the closure, and injecting some energy into the mechanical system. Finally, the folds slam shut and the cycle is complete. It is this rapid closure and the resultant discontinuity in the airflow derivatives that is the primary point of excitation during a normal glottal cycle. This closure instant is referred to as the *epoch.* It can be seen that the normal $U(t)$ will have a power spectral density $S(f)$ asymptotically proportional to $1/f^2$. Since the approximate effect of radiation at the mouth is that of a differentiator, the spectrum of the far-field pressure signal, $U'(t)$, decreases as $1/f$.

The detailed geometry of the vocal folds is somewhat speaker-specific. Additionally, the mechanical properties of the mucous membrane covering the folds are subject to wide variation depending on the speaker's health and ambient conditions. Factors such as these, coupled with the types of voluntary controls mentioned above,

lead to great variety in the shape of $U(t)$. The vocal folds are bathed in mucus which gets somewhat redistributed with each succeding vibratory cycle. In addition, the two folds comprising the glottis are semi-independent mechanical oscillators, and as such, can exhibit varying degrees of chaotic behavior [7]. The net result is often significant cycle-to-cycle variation in the period and shape of $U(t)$. Successful F0 estimators must cope with this variation.

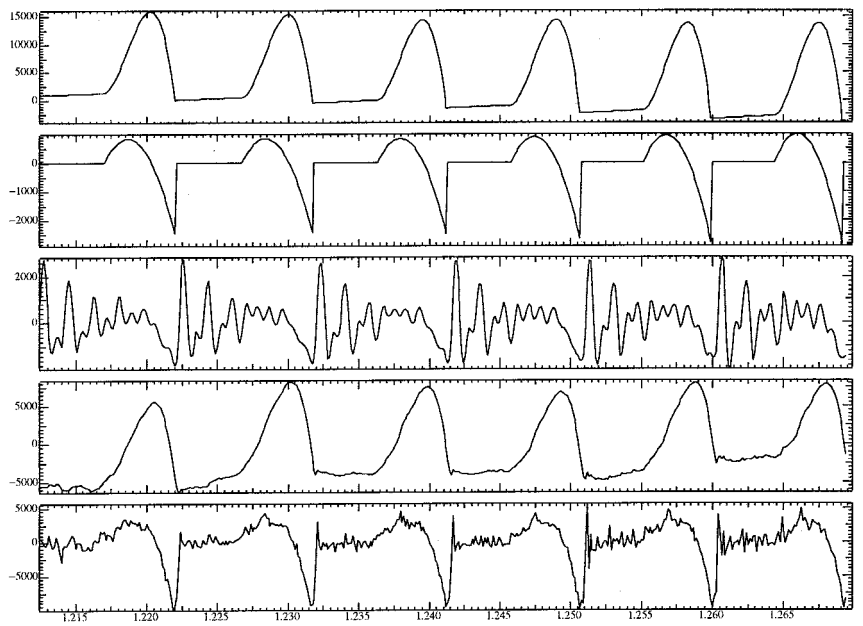

Figure 1. Glottal flow and its first derivative during normal voicing. The horizontal axes display time in seconds. From top to bottom, the signals shown here are synthesized glottal flow $U(t)$; its first derivative, $U'(t)$; a segment of natural speech during production of the phone /*a*/; glottal flow estimated from the natural speech by inverse filtering; and the derivative of the estimated natural flow.

Unvoiced speech is produced when the airflow is forced through a vocal-tract constriction with sufficient velocity to generate significant turbulence. The long-term spectrum of turbulent airflow tends to be a weak function of frequency and is usually approximated by Gaussian noise. Constrictions can occur at several places along the vocal tract from the glottis to the lips. Some speech sounds are produced by complete stoppage of airflow followed by a sudden release, producing an impulsive excitation often followed by a more protracted turbulent excitation. In general, unvoiced speech exhibits little or no periodicity, though occasionally the vocal-tract filtering yields unvoiced signals with considerable periodicity at the resonant frequencies. Some speech sounds are produced by introducing super-glottal vocal-tract

constrictions while voicing is in progress producing *mixed excitation*. Additionally, extremely breathy speech may derive most of its excitation energy from turbulent noise, rather than periodic glottal flow modulation.

1.3. What's the problem?

F0 estimators must cope with mixed excitation. For some applications they must determine the presence or absence of glottis-induced periodicity. The latter determination is referred to as the *voicing classification*. Given the nature of the speech signal, it should now be clear that a whole range of excitation types from "purely voiced" to "purely unvoiced" is possible. However, in this chapter we make the simplifying assumption that only the two extremes exist. While this is patently false, it turns out to offer considerable utility. Many successful speech analysis/synthesis systems are based on this simplifying assumption. RAPT estimates the F0 and the voicing state simultaneously as suggested by Secrest and Doddington [8, 9].

It may be helpful at this point to summarize phenomena observed in natural speech that make F0 difficult to estimate and voicing state difficult to classify.

- F0 changes with time, often with each glottal period.
- Sub-harmonics of F0 often appear that are sub-multiples of the "true" F0.
- In many cases when strong sub-harmonics are present, the most reasonable objective F0 estimate is clearly at odds with the auditory percept.
- Vocal-tract resonances and transmission-channel filtering can emphasize harmonics other than the first, causing F0 estimates that are multiples of the true F0.
- Occasionally F0 actually does jump up or down by an octave!
- Voicing is often very irregular at voice onset and offset leading to minimal waveshape similarity in adjacent periods.
- Panels of expert humans do not agree completely on the locations of voice onset and offset.
- Narrow-band filtering of unvoiced excitation by certain vocal-tract configurations can lead to signals with significant apparent periodicity.
- The amplitude of voiced speech has a wide dynamic range from low in voiced stop consonant closures to high in open vowels.
- It is difficult to distinguish periodic background noise from breathy voiced speech.
- Some voiced speech intervals are only a few glottal cycles in extent.

It follows that considerable temporal context may be required to determine the voicing state and, if voiced, the F0 in human speech. Furthermore, there is no hope of ever inventing an all-purpose algorithm that is correct 100% of the time!

1.4. Whither pitch tracking?

Although some waveform coding techniques require no explicit knowledge of either F0 or the voicing state, many linear-prediction analysis-by-synthesis speech coders

gain in quality and reduced bit rate through pitch determination [10, 11]. Narrow-band coding techniques still benefit significantly from the "vocoder" model, where the type of excitation is chosen on the basis of voicing determination and the estimated F0. Furthermore, knowing the excitation type permits changes in the coding paradigm that offer further economies in bit rate [12]. Although the F0 estimator described below has a delay that would tend to disqualify it from use in common telephony, it is suitable for specialized coders that can tolerate delays on the order of 50 ms and for "off line" analysis situations, such as preparation of speech segment inventories for concatenative text-to-speech systems, compact storage of educational speech material or archival storage of voice mail.

Concatenative speech synthesizers that use a source-filter model for parameter storage, such as that described in chapter 17, require the best possible separation of the excitation source from the vocal-tract filter. A clean separation along these lines during analysis permits increased flexibility in changing F0 and voice quality during resynthesis, where the segmental F0 and voice quality adjustments can be substantially different from those in the source utterances for the concatenated elements. By weighting consistently the closed-glottis interval of each period more heavily than the open glottis interval, the acoustic contributions due to the details of airflow through the glottis and the variations in vocal-tract resonant frequencies and bandwidths due to the glottal leakage are reduced. This differential weighting can be achieved consistently if the time locations of the glottal epochs are known [13]. Consistent positioning of the analysis window relative to the epochs also allows the use of smaller windows, resulting in less smoothing of the time-varying vocal-tract filter estimates. This results in more clearly articulated synthesis.

As automatic speech recognizers are incorporated into speech understanding systems, analysis of utterance prosody will become increasingly important for detecting emphasis and disambiguating meaning [14–17]. Since F0 variation is an important component of prosodic implementation, reliable F0 determination will be required in thse systems. Basic and applied studies of speech prosody currently under way already demand reliable estimates of F0 over large corpora of speech data [18].

Speech processing aids for the hearing impaired often translate the voicing state and F0 into the tactile or visual sense domains, or recode it for presentation in cochlear implants. Transformations of the F0 and voicing information into an alternate auditory signal [19] or into the tactile domain [20–22] show considerable promise as communication aids for the hearing impaired.

2. Approaches to F0 estimation

F0 estimators often have three major components: *i)* A pre-processing, or signal-conditioning stage, *ii)* a generator of candidate estimates for the true period sought and *iii)* a "post-processing" stage that selects the best candidate and refines the F0 estimate. In this section, some commonly used approaches to these operations are outlined.

2.1. Pre-processing

The choice of pre-processing depends to some extent on the nature of the candidate generator that is to follow. The aim of pre-processing is to remove interfering signal components, such as extraneous noise, vocal-tract filter influences, DC offset, etc. and to transform the signal to better match the expectations of later stages.

It has been suggested that low-pass filtering improves F0 estimation performance because it removes an apparent loss of periodicity in the voiced speech spectrum at higher frequencies [1]. While this aperiodicity is certainly observable in the short-time spectrum, it is due to an undetermined mix of two effects: *i)* averaging periods of different length within the spectral analysis window, and *ii)* a predominance of random over "periodic" excitation in those spectral regions. Of course, if the candidate generation stage is also averaged over several periods, then the low-pass filtering will have a beneficial (or at least non-harmful) effect regardless of the explanation for the periodicity loss at high frequencies. Since RAPT does not average over several periods, low-pass filtering is unnecessary. Since bandwidth reduction in general tends to increase inter-sample correlation, it can have a detrimental effect in systems which rely on correlation values as an indicator of periodicity (and hence, of voicing state).

Non-linear operations on the speech signal such as cubing and center/peak clipping have been shown to have the effect of flattening the spectrum of the signal passed to the candidate generator [23, 24]. This has the effect of increasing the distinctiveness of the true period peaks in autocorrelation functions, as described below. However, it also has the effect of destroying some information that may lead to better estimates of period length in candidate generators that do not average over more than one period.

Auto-regressive (linear predictor) inverse filtering has also been suggested as a pre-processing step to flatten the signal spectrum [25, 9, 26]. When the aim is to detect glottal epochs, inverse filtering is extremely beneficial [27–29, 13]. In other contexts its benefits are less clear and it has been found to degrade the voiced/unvoiced decision and F0 estimation in some cases. When the speech signal is rich in harmonics of F0, inverse filtering has the beneficial effect of removing the contamination by the vocal-tract resonances. However, when only a few harmonics are present, for instance in voiced obstruents, breathy speech, or falsetto, inverse filtering can remove all traces of the F0 one is trying to estimate!

Most candidate generators described below are impaired by the presence of significant DC or very-low-frequency components in the speech signal. Thus, some form of high-pass filtering is recommended for systems designed to be maximally robust. Of course, it is desirable (but not essential for many candidate generators) for the filter to pass all voice harmonics including the first.

2.2. Period candidate generating functions

Period candidate generators attempt to capitalize on the following characteristics of voiced speech:

- The glottal period usually varies by only a small percentage from one period to the next.
- The vocal tract filter varies slowly in comparison with the glottal inter-pulse interval. Hence, adjacent periods of the speech signal tend to have similar shapes.

(Note the distinction between the determination of glottal pulse location which is phase sensitive, and period determination which is not.)

2.2.1. Direct waveform processors

The speech waveform and various linear and nonlinear filterings of it have been subjected to peak/valley detectors and inter-period similarity detectors. These algorithms estimate the period by searching for similarity in the pattern of gross speech waveform features from one period to the next. The algorithm described in [30] is probably the most widely used example of this. This algorithm performs well in high SNR conditions and is of low computational complexity. It is capable also of detecting local period variations, since it does not integrate the final estimate over more than a single period. Where computational resources are at a premium, this algorithm is a good choice.

2.2.2. Autocorrelation

The autocorrelation function (ACF) of the speech signal, or of a pre-processed version of it, is a traditional source of period candidates [31]. Given s_p, $p = 0, 1, 2, \ldots$, a sampled speech signal with sampling interval $T = 1/F_s$, analysis frame interval t, and analysis window size w, at each frame we advance $z = t/T$ samples with $n = w/T$ samples in the autocorrelation window. w is chosen to be at least twice the longest expected glottal period; s is assumed to be zero outside the window. t is sized to sample adequately the time course of changes in F0. The ACF of K samples length, $K < n$, may then be defined as

$$R_{i,k} = \sum_{j=m}^{m+n-k-1} s_j s_{j+k}, \quad k = 0, K-1;\ m = iz;\ i = 0, M-1, \tag{2.1}$$

where i is the frame index for M frames, and k is the *lag index* or *lag*. As outlined in [24], several linear and linear operations have been applied to the speech signal with the aim of flattening its spectrum and thus causing R to more closely approximate a sequence of impulses with significant amplitude only at the "true" period multiples. Another outcome of pre-processing can be to reduce the computational cost by very coarsely quantizing the speech signal [32].

While autocorrelation has performed well in many contexts and is relatively noise immune, it has two flaws that reduce its utility as a period candidate generator. The chief disadvantage is the relatively large time window over which the autocorrelation must be computed in order to cover adequately F0 ranges encountered in human speech. This precludes resolution of cycle-to-cycle variation in shorter periods. Rapid F0 changes can result in the loss of a clear peak in R at any "true" period. A second difficulty is that the statistical significance (noise immunity) of the peak estimates vary as a function of the lag index, k, since the summation interval shrinks as k increases. Thus, in order to maintain significance at the longest lags (lowest F0), the window is excessively large at the shorter lags.

2.2.3. Cepstrum

Related to the ACF is the *cepstrum* as described originally in [33] and applied to F0 estimation in [34]. The cepstrum is defined as the inverse Fourier transform of the short-time log magnitude spectrum. Given s_m, i, z, and n as defined above, the short-time log-magnitude spectrum computed as,

$$S_{i,p} = \log || \sum_{r=0}^{n-1} s_{r+m} e^{-j2\pi rp/n} ||, \quad p = 0, n-1; \; m = iz; \; i = 0, M-1, \tag{2.2}$$

has maxima at values of p corresponding to integer multiples of $nF0/F_s$, provided the signal shows periodicity in the window of length n. The cepstrum $c_{i,k}$ for frame i and lag k (or *quefrency* k, to use the terminology of Bogart, et al) may then be computed as

$$c_{i,k} = \frac{1}{n} || \sum_{r=0}^{n-1} S_{i,r} e^{j2\pi rk/n} ||. \tag{2.3}$$

The cepstrum tends to have local maxima at times, kT, corresponding to integer multiples of the glottal period. The log operator on the speech magnitude spectrum tends to flatten the harmonic peaks in the spectrum and thus lead to the more distinct period peaks in the cepstrum. Unfortunately, the interval of speech over which the spectrum, and hence the cepstrum, must be computed is the same as that required for the ACF, and thus, the cepstrum shares the disadvantages of the ACF. In other respects, it seems to have similar usefulness as a candidate generator.

2.2.4. Cross-correlation

Some of the shortcomings in the ACF are overcome by using the cross-correlation function (CCF), χ. Here, w can be chosen to be on the order of a single average

glottal period. With s, M, K, i, and n as defined above, χ, is defined as

$$\chi_{i,k} = \sum_{j=m}^{m+n-1} s_j s_{j+k}, \quad k = 0, K-1;\ m = iz;\ i = 0, M-1. \tag{2.4}$$

Thus, the correlation interval w can be chosen independently of the interval being searched for period candidates. Note that there are no theoretical bounds on the value of χ, and in fact, $\chi_{i,k}$ for $k = 0$ may be smaller than for some other value of k, unlike $R_{i,0}$ which always has its largest value for $k = 0$. Thus, when s is changing rapidly in amplitude, normalizing $\chi_{i,k}$ by $\chi_{i,0}$ is not sufficient to permit reliable candidate choices using simple threshold logic.

2.2.5. AMDF

When computational cost was more of an issue and it was common that vector differences could be computed significantly faster than dot products, the comb filter or average magnitude difference function (AMDF) was of interest [35, 36]. It is included here primarily for its historical value, though there may still be some contexts where its computational characteristics could pay off.

Given s, w, M, K, i, and n as defined above, the AMDF, D, is defined as

$$D_{i,k} = \sum_{j=m}^{m+n-1} |s_j - s_{j+k}|, \quad k = 0, K-1;\ m = iz;\ i = 0, M-1. \tag{2.5}$$

The near-zero *minima* in D become the period candidates for the post-processing stage. Again, the window size w need only be on the order of an average glottal period, and can remain constant over all values of k. If the amplitude s is changing rapidly, then the local minimum in D at the "true" period can be significantly above zero, making correct period determination especially difficult. Nonetheless, the AMDF has been used extensively and is, in fact, the candidate generator in the U.S. Government standard LPC-10 vocoder [37].

2.2.6. Normalized cross-correlation

The RAPT algorithm to be described in section 3 is based on the normalized cross-correlation function (NCCF) [38]. The NCCF overcomes all of the shortcomings of the other candidate generators described above at a slight increase in computational complexity. Let s_m be a non-zero sampled speech signal with zero mean, and let w, M, K, and n be as defined above. Once again, w is chosen to be in the neighborhood of the expected F0 period. The NCCF, $\phi_{i,k}$ at lag k and analysis frame i is

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}}, \quad k = 0, K-1;\ m = iz;\ i = 0, M-1, \tag{2.6}$$

where

$$e_j = \sum_{l=j}^{j+n-1} s_l^2. \tag{2.7}$$

Note that $-1.0 \leq \phi \leq 1.0$. $\phi_{i,k}$ tends to be close to 1.0 for lags corresponding to integer multiples of the "true" period, regardless of rapid changes in the amplitude of s, provided that the shapes of successive periods are similar. The correlation interval w may be chosen independently of the full F0 range under consideration. For s_m that is white noise, $\phi_{i,0} = 1.0$ and $\phi_{i,k}$ approaches zero for $k \neq 0$ as w increases. For practical values of w it is still true that the NCCF of noise at all non-zero lags has a magnitude considerably less than 1. These properties of the NCCF are independent of the amplitude of s.

We can represent $\phi_{i,k}$ graphically by assigning lag to the ordinate, frame index (or time) to the abscissa, and the value of ϕ at the corresponding time and lag to the degree of shading, with dark shading representing high values (close to 1.0) and white representing low values (close to -1.0). These graphical representations are referred to as *correllograms*.

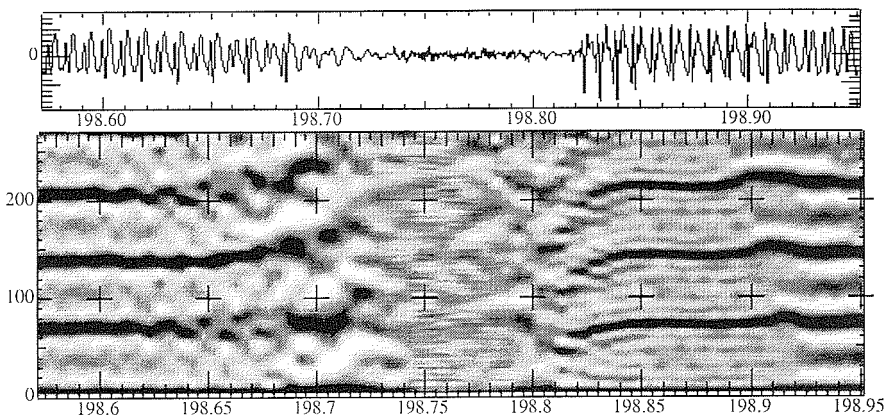

Figure 2. Speech signal and correllogram of the utterance fragment "in San [Diego]" produced by a male talker. The horizontal axis is time in seconds; the vertical is lag number in samples ($F_s = 8kHz$). Correlation values close to 1.0 are the darkest; those close to -1.0, the lightest. Note the generally low non-zero-lag correlation values in the /s/ region around 198.75 sec. Voiced regions exhibit dark horizontal striations corresponding to peaks in the NCCF at lags that are multiples of the fundamental period. In this fragment, the "true" period corresponds to the striation near lag 65. Between the times of 198.625 and 198.7 the correlation peak at twice the correct period is stronger and more consistent than the "true" peak.

An utterance containing clear and problematic voiced speech segments and an unvoiced segment is shown in fig. 2. The only *local* evidence for the true period is

the location and height of maxima in the NCCF at each frame. The only hope for choosing the correct peak candidate in the vicinity of 198.65 sec. in this utterance is to consider the candidate peaks in a large temporal context. Note that, in general, the NCCF of voiced speech has maxima with comparable amplitudes at lag intervals corresponding to integer multiples of the fundamental period while the NCCF of unvoiced speech has its most prominent maximum at zero lag.

2.3. Post-processing

Many post-processing techniques have been developed to cope with the difficulties summarized in section 1.3. Of course, the most primitive of these is simply the selection of the "most likely" period candidate, basing the decision solely on time-local evidence, such as peak or valley amplitude.

More evolved techniques include various heuristics that examine the current F0 hypotheses in relation to past F0 estimates. In some cases, the frequency interval searched for F0 is restricted to the neighborhood of past "reliable" estimates.

A straightforward and often successful post-processing strategy is the use of median smoothing [39]. This has the desirable property of ignoring isolated outliers while preserving both the fine-grained variations in F0 and the sharpness of true step transitions.

Among the most successful techniques is dynamic programming (DP). Probably the earliest use of DP in F0 estimation was reported in [40]. Later, it was clearly outlined in [41, 42] how DP can be applied to the joint problem of estimating and smoothing speech parameters, including F0. Secrest & Doddington [8, 9] described the use of DP to integrate the voicing decision with F0 estimation and convincingly demonstrated improved performance on both.

Sufficiently precise estimation of F0 from a sampled speech signal may require refinement of the period estimate after a "final" candidate selection has been performed. This may be done by effectively increasing the sampling rate of the generating function in the vicinity of the peak, and then relocating the peak at the higher sample rate through band-limited interpolation [26]. Alternately, a polynomial fit to the coarsely-sampled generating function in the vicinity of the peak can be evaluated to determine the point of zero first derivative, thus yielding the virtual peak location with high precision.

3. RAPT

The primary aim in the development of the F0 estimator described here was to obtain the most robust and accurate estimates possible, with little thought to computational complexity, memory requirements or inherent processing delay. However, it will be seen that several efficiency enhancements have been incorporated that significantly reduce computational cost while maintaining the desired accuracy. Although the delay inherent in RAPT probably disqualifies if from use in

standard telephony, it does operate continuously and can be used anywhere a delay of a few tens of milliseconds can be tolerated. RAPT is designed to work at any sampling frequency and frame rate over a wide range of possible F0, speaker and noise conditions. Parameter adjustments permit adaptation of the algorithm for speed/accuracy tradeoffs and to match peculiar voice or recording conditions.

The following characteristics of typical speech signals and of their NCCFs are exploited in the algorithm:

- The local maximum in ϕ corresponding to the "true" F0 for voiced speech (excepting the maximum at zero lag) is usually the largest and is close to 1.0.
- When multiple maxima in ϕ exist and have values close to 1.0, the maximum corresponding to the shortest period is usually the correct choice.
- True ϕ maxima in temporally adjacent analysis frames are located usually at comparable lags, since F0 is a slowly-varying function of time.
- The "true" F0 occasionally changes abruptly by doubling or halving.
- Voicing tends to change states with low frequency.
- The largest non-zero-lag maximum in ϕ for unvoiced speech is usually considerably less than 1.0.
- The short-time spectra of voiced and unvoiced speech frames are usually quite different.
- Amplitude tends to increase at the onset of voicing and to decrease at offset.

3.1. Algorithm outline

Here is an overview of the steps that constitute RAPT:

- Provide two versions of the sampled speech data; one at the original sample rate; another at a significantly reduced rate.
- Periodically compute the NCCF of the low sample rate signal for all lags in the F0 range of interest. Record the locations of local maxima in this first-pass NCCF.
- Compute the NCCF of the high sample-rate signal only in the vicinity of promising peaks found in the first pass. Search again for local maxima in this refined NCCF to obtain improved peak location and amplitude estimates.
- Each peak retained from the high-resolution NCCF generates a candidate F0 for that frame. At each frame the hypothesis that the frame is unvoiced is also advanced.
- Dynamic programming is used to select the set of NCCF peaks or unvoiced hypotheses across all frames that best match the characteristics mentioned above.

Though inspired by and similar to the "Integrated Pitch Tracker" [9], RAPT differs in several details. Significant differences are:

- The NCCF is computed on the speech signal, rather than the LPC residual.
- Two stages of NCCF are used to reduce the overall computational load. This is similar in some respects to SIFT, but uses peak interpolation at the original sample rate for increased accuracy.

– The documentation included below is intended to provide sufficient detail to reproduce exactly the results obtained using the "get_f0" program in the *waves+* software package from Entropic Research Laboratory, Inc.

For convenience, all symbols and constants relevant to the algorithm description are defined here. The numerical values of the constants provided here were determined by hill climbing using a hand-marked speech database composed of several adult male and female talkers.

Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20

Symbol	Meaning
x_m	m^{th} sample of the input speech signal
F_s	sample rate of speech signal $= 1/T$
F_{ds}	reduced sample rate of speech for first-pass NCCF
$round(v)$	the integer that is closest to v
n	the number of samples correlated at each lag $= round(wF_s)$
z	the frame step size in samples $= round(tF_s)$
i	the analysis frame index incrementing at a rate of $1/Tz$
K	the longest lag at each frame $= round(F_s/F0_{min})$
$\phi_{i,k}$	normalized cross-correlation for frame i at lag k

3.2. Preprocessing

The algorithm described here does not require, inherently, any special preprocessing of the input speech signal, and it performs well on speech sampled at any typical audio sampling rate (6 kHz $\leq F_s \leq$ 44 kHz). However, the cost of computation grows roughly linearly with F_S, so in some cases it may be economical to down-sample the speech in preparation. Where the background noise in the signals to be processed has a significant periodic component, attempts should be made to remove the periodicity. While F0 estimation is only weakly affected by modest levels of periodic noise, the voicing-state determination can be strongly affected. Possible approaches include using an inverse filter trained on the periodic noise, or a comb filter tuned to cancel the (fixed) harmonic spectrum of the noise (e.g. 50/60 Hz line-induced hum/buzz). In extreme cases of periodic noise in the background, the reliability of the voicing determination can be improved through the use of center clipping, possibly combined with the addition of white noise at a level sufficient to mask the background periodicity, but several dB below the typical voiced-speech amplitude.

3.3. Two-pass NCCF

The NCCF described in section 2.2.6 is the source for period candidates. Its computation is the dominant cost in the algorithm. One device for reducing the computation is to limit the range of F0 values searched. However, a general-purpose F0 estimator should search at least the range, $50 \leq F0 \leq 500$.

Since both n and K grow as F_s, the cost of the NCCF grows as ${F_s}^2$. This is dealt with using a two-pass procedure that has a computational cost roughly proportional to F_s. In the first pass, the input signal is resampled at a lower rate, F_{ds}, determined by

$$F_{ds} = \frac{F_s}{round(F_s/(4F0_{max}))}. \tag{3.1}$$

The low-pass filter applied before decimation is a symmetric FIR obtained by truncating the impulse response of an ideal $F_{ds}/2$ low-pass filter with a 5 ms duration Hanning window. The NCCF of the down-sampled signal is computed at all lags, k, $F_{ds}/F0_{max} \leq k \leq K$. The maximum value of ϕ in this range, ϕ_{max}, is noted. All local maxima in ϕ that exceed (CAND_TR $\times$ ϕ_{max}) are marked. More precise estimates of correlation peak location and amplitude at F_s are then obtained using parabolic interpolation on the three samples of ϕ defining each peak at F_{ds}. If the number of these peaks exceeds (N_CANDS $-$ 1), they are sorted by decreasing amplitude, and the top (N_CANDS $-$ 1) are saved.

In the second pass, ϕ is computed using the original speech signal at F_s, but only for seven lags in the vicinity of each refined peak estimate kept from the first pass. A new ϕ_{max} is also found. ϕ is assumed zero at those lags not computed. All

peaks in this higher resolution ϕ exceeding (CAND_TR$\times\phi_{max}$) are marked. In both passes (CAND_TR $\times$ ϕ_{max}) is used as the peak screening level, rather than simply CAND_TR to provide some normalization of possibly reduced peak value due to additive noise in a truly voiced signal. Again, if the number of maxima exceeds (N_CANDS – 1), they are sorted and only the top (N_CANDS – 1) are kept. The higher resolution peaks are not further refined using parabolic interpolation at this time.

3.4. Notes on computing the NCCF

If the speech signal has non-zero mean in the correlation window w, or if there is very low frequency noise present, ϕ from eq. (2.6) can yield high correlation at all lags in the range searched for F0. This is especially troublesome when "silent" intervals, or low-amplitude unvoiced intervals are to be classified as voiced or unvoiced largely on the basis of the amplitude of ϕ. The solution to this used in RAPT is to subtract the local mean in each reference window from all samples involved in the computation at each frame. The NCCF is then computed on this modified signal segment. If x_m, $m = iz, iz+1, iz+2, \ldots$ is the non-zero-mean input signal for frame i, then the signal, $s_{i,j}$, to be passed to the correlator at frame i is

$$s_{i,j} = x_{m+j} - \mu_i, \quad m = iz;\ j = 0, n+K-1, \tag{3.2}$$

where

$$\mu_i = \frac{1}{n}\sum_{j=m}^{m+n-1} x_j. \tag{3.3}$$

At first glance it may appear that the energy normalization term in the denominator of eq. (2.6) would double the number of arithmetic operations for the NCCF as compared to the CCF. However, note that e_m is only computed once for all lags at a given frame and that e_{m+k} can be modified incrementally by subtracting s^2_{m+k-1} and adding $s^2_{m+k+n-1}$ at each lag. Thus, the increase in computational cost of the NCCF over the CCF is minimal. However, the precision required for this incremental adjustment of e_{m+k} is considerable, since the magnitude of the increments and of the sum of squares can differ greatly. To cope with this, double-precision arithmetic should be used for e_{m+k} and it should be limited to values greater than some positive minimum (e.g. 1) to cope with numerical imprecision.

Analysis frames which contain no energy in the correlation reference window or which have no local maxima for some other reason, yield no period candidates. In these cases, the highest value of ϕ reported for the frame is zero and the frame is classified as unvoiced.

The signal in "silent" regions of even carefully digitized speech material often contains a significant periodic component. In these cases, if ϕ is computed as in eq. (2.6), high correlations may be observed that can lead to an incorrect voicing

decision. Thus, it is useful to incorporate some knowledge of absolute signal level. In the second-pass NCCF, we do this through an additive constant, A_FACT, in the denominator term of ϕ. The modified NCCF at lag k and frame i is now defined as

$$\phi_{i,k} = \frac{\sum_{j=0}^{n-1} s_{i,j} s_{i,j+k}}{\sqrt{\text{A\_FACT} + e_0 e_k}}, \tag{3.4}$$

where

$$e_j = \sum_{l=j}^{j+n-1} s_{i,l}^2, \tag{3.5}$$

and $s_{i,l}$ is defined in eq. (3.2). The first-pass NCCF uses the same formula, but *without* A_FACT, and k ranges from $round(F_{ds}/F0_{max})$ to $K-1$.

3.5. Post-processing with dynamic programming

Dynamic programming [43, 44] is now applied to select the best F0 and voicing state candidates at each frame based on a combination of local and contextual evidence. Although the entire utterance is available for this optimization, the solution typically converges in a few tens of milliseconds. For more background on the approach described below, the reader is referred to [41, 42].

Let I_i be the number of states proposed at frame i, which is one plus the number of non-zero-lag local maxima selected from ϕ at frame i ($1 \leq I_i \leq$ N_CANDS). Thus, at each frame, $I_i - 1$ possible fundamental frequencies (voiced states) and one unvoiced state will be proposed. Also, let $C_{i,j}$ be the value of the j^{th} local maximum in ϕ at frame i. These are the retained peak values from the second-pass NCCF. Finally, let $L_{i,j}$ be the sample lag at which $C_{i,j}$ was observed.

We may now define an objective function as the *local cost* for proposing that frame i is voiced with period $TL_{i,j}$ as

$$d_{i,j} = 1 - C_{i,j}(1 - \beta L_{i,j}), \; 1 \leq j < I_i, \tag{3.6}$$

while the cost for the single unvoiced hypothesis at frame i is

$$d_{i,I_i} = \text{VO\_BIAS} + \max_j(C_{i,j}) \tag{3.7}$$

$\beta = \text{LAG\_WT}/(F_s/F0_{min})$. LAG_WT permits adjustment of the degree to which correlations at longer lags are devalued so as to encourage the selection of shorter periods. This local cost function favors $C_{i,j}$ close to 1.0 and shorter lags for voiced frames, and $C_{i,j}$ close to zero for unvoiced frames. The bias term VO_BIAS permits adjustment of the likelihood of a voiced decision.

The inter-frame F0 transition cost δ at frame i when hypotheses j and k at the current and previous frames are both voiced is defined as

$$\delta_{i,j,k} = \text{FREQ\_WT} \times \min\{\xi_{j,k}, (\text{DOUBL\_C} + |\xi_{j,k} - \ln(2.0)|)\}, \tag{3.8}$$

where

$$\xi_{j,k} = |\ln \frac{L_{i,j}}{L_{i-1,k}}|, \quad 1 \leq j < I_i; \; 1 \leq k < I_{i-1} \tag{3.9}$$

and DOUBL_C is a positive constant. This makes the transition cost an increasing function of inter-frame proportional frequency change, but allows octave jumps at some specifiable cost. FREQ_WT is a positive constant that adjusts the cost of inter-frame F0 changes. DOUBL_C adjusts the cost of an exact octave increase or decrease in F0.

When the current and previous frames are both proposed as unvoiced:

$$\delta_{i,I_i,I_{i-1}} = 0. \tag{3.10}$$

The inter-frame transition cost applied when the voicing states proposed for the previous and current frames differ is computed as

voiced-to-unvoiced:

$$\delta_{i,I_i,k} = \text{VTRAN\_C} + (\text{VTR\_S\_C})S_i + (\text{VTR\_A\_C})rr_i, \quad 1 \leq k < I_{i-1}; \tag{3.11}$$

unvoiced-to-voiced:

$$\delta_{i,j,I_{i-1}} = \text{VTRAN\_C} + (\text{VTR\_S\_C})S_i + \text{VTR\_A\_C}/rr_i, \quad 1 \leq j < I_i, \tag{3.12}$$

where VTRAN_C, VTR_S_C and VTR_A_C are positive constants, S_i is a spectral *stationarity* function, and

$$rr_i = \frac{rms(i,h)}{rms(i-1,-h)} \tag{3.13}$$

is the RMS ratio across the proposed voicing boundary.

$$rms(i,h) = \sqrt{\frac{\sum_{j=0}^{J-1}(W_j s_{j+m+h})^2}{J}}, \; m = iz \tag{3.14}$$

W is a Hanning window of length $J = .03F_s$; z is the frame step; h is an offset that adjusts the window centers for the current and previous *rms* measurements to be

20 ms apart, regardless of the frame step size, z. If the speech signal amplitude is increasing, $rr > 1$, if it is decreasing, $0 < rr < 1$.

S is an inverse function of the Itakura distortion [45] measured across the proposed voicing boundary.

$$S_i = \frac{0.2}{itakura(i, i-1) - 0.8}, \tag{3.15}$$

where the spectral distortion, $itakura(i, i-1)$, is computed using Hanning windows sized and positioned as above for the RMS ratio. The order of the LPC analysis, O is chosen according to

$$O = 2 + round(F_s/1000); \tag{3.16}$$

the signal is preemphasized using a 1st-order filter coefficient of e^{-7000/F_s}; and the autocorrelation LPC method is used on the speech signal sampled at F_s.

These voicing-state transition costs decrease when the signal spectrum is changing rapidly as it does across voicing class boundaries, and when the signal amplitude change is in accord with expectations at voice onset (rising) and offset (falling). VTRAN_C provides a fixed penalty for voicing state change, regardless of the changes in the speech signal, to encourage estimation behavior in line with the general observation that speech changes voicing state relatively infrequently.

We may now define the recursion for the optimal objective function for frame i as

$$D_{i,j} = d_{i,j} + \min_{k \in I_{i-1}} \{D_{i-1,k} + \delta_{i,j,k}\}, \quad 1 \leq j \leq I_i, \tag{3.17}$$

with the initial conditions

$$D_{0,j} = 0, \quad 1 \leq j \leq I_0; \quad I_0 = 2. \tag{3.18}$$

For each state at each frame we save the "back pointers"

$$q_{i,j} = k_{min}, \tag{3.19}$$

where k_{min} at each frame are the indices, k, which minimize $D_{i,j}$, so that the optimal state sequence can be retrieved. Back pointers from each state at frame i may be traced backwards until they converge to a common, globally optimal state at frame $i - l$, where lt is the latency of the decision. In practice, this latency for the F0 estimate is rarely greater than 100 ms. Thus, it is feasible to implement F0 estimators using this algorithm that can operate continuously, in real time, with modest delay. The coarse F0 estimate for the frame is

$$F0_i = \frac{F_s}{L_{i,j}}, \tag{3.20}$$

where the values of j are those which result in the global minimum value for D. This estimate is refined finally using a parabolic fit to the three points in ϕ comprising the peak. The point where the first derivative of the fit is zero is taken as the "true" peak. When F_s is only a few times $F0_{max}$, or where extreme precision in the F0 estimates is required, it may be advisable to interpolate using a low-pass filter sampled at a rate high enough to yield the required precision [26].

The algorithm described above has been used with satisfactory results on speech recordings varying in quality from noisy telephone to quiet laboratory conditions. A C-language implementation on a MIPS R3000 based UNIX workstation runs continuously in less than half real time on speech sampled at 8kHz with a frame rate of 100 Hz. Thus, "real-time" systems, such as modest delay vocoders and speech parameter displays for speech training can be built around this F0 estimator. The algorithm has been embedded in a commercially available speech-processing package and is in widespread use in speech research laboratories.

4. Discussion and summary

This chapter has provided a definition of the "pitch tracking" problem, and through the complete description of RAPT, an operational definition of the voice fundamental frequency, F0, which is a strong physical correlate of the psychological percept of pitch. As a by-product of integrating all available information, RAPT makes a binary voicing classification regarding the presence or absence of voicing in the speech signal. Within the limited model of two-state voicing and single-frequency voiced excitation, RAPT provides reliable F0 and voicing estimates by considering all possibilities simultaneously in a large temporal context.

For some applications a graded measure of voicing, rather than a two-state classification, may be required to quantify the mix of periodic and random components in the speech signal. The maximum non-zero-lag value in the NCCF of eq. (3.4) is one such measure. Once F0 has been estimated using RAPT, the inter-period correlation over exactly one glottal period can be computed by setting w to the estimated period and k to the lag corresponding to the estimated F0. The periodic/aperiodic ratio also may be estimated as a function of frequency, by computing the NCCF on band-passed versions of the speech signal.

This chapter provided an overview of the speech production process that should help in the understanding of the pitch and voicing determination problems. The characteristics of the normal speech signal that make F0 analysis difficult were outlined, and the limitations of the speech-production model were described. Some applications of F0 analysis were then presented. Some commonly-encountered F0 candidate generating functions were defined. Finally, a widely used F0 and voicing determination algorithm, RAPT, was described. RAPT owes a significant debt to much work that has gone before, but the author would especially like to acknowledge the contributions of Bishnu Atal [38], John Markel [26], Herman Ney [41, 42], Bruce Secrest and George Doddington [9].

References

[1] W. Hess, *Pitch Determination of Speech Signals.* Berlin, Heidelberg, New York: Springer-Verlag, 1983.

[2] W. Hess, "Pitch and voicing determination," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, eds.), New York: Marcel Dekker, 1991.

[3] J. L. Flanagan, *Speech Analysis Synthesis and perception.* Berlin, Heidelberg, New York: Springer-Verlag, 1972.

[4] I. R. Titze, *Principles of Voice Production.* Englewood Cliffs, NJ: Prentice-Hall, 1994.

[5] G. Fant and Q. Lin, "Frequency domain interpretation and derivation of glottal flow parameters," Tech. Rep. QPSR 2-3, Speech Transmission Lab., Royal Inst. Technology, Stockholm, Sweden, 1988.

[6] J. Pierrehumbert and D. Talkin, "Lenition of /h/ in glottal stop," in *Papers in Laboratory Phonology II* (G. J. Docherty and R. D. Ladd, eds.), (Cambridge, UK), Cambridge University Press, 1992.

[7] I. R. Titze, Herzel, and Baken, "Chapter 4," in *Vocal Fold Physiology: Frontiers in Basic Science*, Singular Publishing, 1993.

[8] B. G. Secrest and G. R. Doddington, "Postprocessing techniques for voice pitch trackers," in *1982 International Conference on Acoustics, Speech and Signal Processing*, pp. 172–175, IEEE, New York, NY, 1982.

[9] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (Boston), pp. 1352–1355, IEEE, New York, NY, 1983.

[10] M. Yong and A. Gersho, "Efficient encoding of the long-term predictor in vector excitation coders," in *Advances in Speech Coding* (B. S. Atal, V. Cuperman, and A. Gersho, eds.), pp. 329–338, Dordrecht, Holland: Kluwer Academic Publishers, 1991.

[11] R. Salami, C. Laflamme, J.-P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (pcs)," *IEEE Trans. Vehic. Techn.*, vol. 43, no. 3, pp. 808–816, 1994.

[12] A. V. McCree and T. P. Barnwell, "Improving the performance of a mixed-excitation LPC vocoder in acoustic noise," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (San Francisco), pp. II137–II138, 1992.

[13] D. Talkin and J. Rowley, "Pitch-synchronous analysis and synthesis for TTS systems," in *Proceedings of the ESCA Workshop on Speech Synthesis* (C. Benoit, ed.), (Gieres, France), Imprimerie des Ecureuils, 1990.

[14] M. Ostendorf, P. Price, J. Bear, and C. W. Wightman, "The use of relative duration in syntactic disambiguation," in *DARPA Speech and Natural Language Workshop*, pp. 26–31, June 1990.

[15] M. Ostendorf, C. W. Wightman, and N. M. Veilleux, "Parse scoring with prosodic information: An analysis/synthesis approach,," *Computer Speech and Language*, pp. 193–210, July 1993.

[16] N. M. Veilleux and M. Ostendorf, "Probabilistic parse scoring with prosodic information," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, pp. II51–II55, IEEE, New York, NY, April 1993.

[17] N. M. Veilleux and M. Ostendorf, "Prosody/parse scoring and its application in atis," in *ARPA Workshop on Human Language Technology*, pp. 335–340, March 1993.

[18] J. F. Pitrelli, M. E. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the tobi framework," in *International Conference on Spoken Language Processing*, (Yokohama), September 1994.

[19] S. M. Rosen, A. J. Fourcin, and B. C. Moore, "Voice pitch as an aid to lipreading," *Nature*, vol. 291, pp. 150–152.

[20] L. Hanin, A. Boothroyd, and T. Hnath-Chisolm, "Tactile presentation of voice fundamental frequency as an aid to speechreading of sentences," *Ear and Hearing*, vol. 9, pp. 329–334.

[21] L. E. Bernstein, S. P. Eberhardt, and M. E. Demorest, "Single-channel vibrotactile supplements to visual perception of intonation and stress," *Acoustical Society of America*, vol. 85, pp. 397–405.

[22] E. T. Auer and L. E. Bernstein, "Sensory substitution to enhance spoken language comprehension by deaf people: Acoustic-to-vibrotactile transformations of voice fundamental frequency," *Acoustical Society of America*, Manuscript submitted for publication.

[23] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio and Electroacoustics*, vol. ASSP-16, pp. 262–266, 1968.

[24] L. R. Rabiner, B. S. Atal, and M. R. Sambur, "LPC prediction error - analysis of its variation with the position of the analysis frame," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 434–442, Oct. 1977.

[25] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Acoustical Society of America*, vol. 50, no. 2, Part 2, pp. 637–655, 1971.

[26] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio and Electroacoustics*, vol. AU-25, pp. 367–377, Dec. 1972.

[27] T. V. Ananthpadmanabha and B. Yegnanaraya, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, August 1979.

[28] M. D. Anderson, "Pitch determination of speech signals," Master's thesis, M. I. T., February 1986.

[29] D. Talkin, "Voice epoch determination with dynamic programming," *Acoustical Society of America*, vol. 85, no. S1, p. S149, 1989.

[30] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *Acoustical Society of America*, vol. 46, pp. 442–448, August 1969.

[31] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 24–33, February 1977.

[32] J. J. Dubnowski, H. L. Shaffer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 2–8, February 1976.

[33] B. P. Bogart, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Symposium on Time Series Analysis* (M. Rosenblatt, ed.), (New York), pp. 209–243, John Wiley and Sons, 1963.

[34] A. M. Noll, "Cepstrum pitch determination," *Acoustical Society of America*, vol. 41, pp. 293–309, February 1967.

[35] J. A. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353–362, Oct. 1974.

[36] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353–362, Oct. 1974.

[37] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology Magazine*, pp. 40–49, April 1982.

[38] B. S. Atal, *Automatic Speaker Recognition Based on Pitch Contours*. PhD thesis, Polytechnic Institute of Brooklyn, 313 N. First St., Ann Arbor, Michigan, 1968.

[39] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Application of nonlinear smoothing algorithm to speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 552–557, 1975.

[40] W. R. Bauer and W. A. Blankenship, "DYPTRACK – a noise-tolerant pitch tracker," Tech. Rep. NASL-S-210, 525, Dept. of Defence (NSA), U. S. A., 1974. Unclassified Report.

[41] H. Ney, "A dynamic programming technique for nonlinear smoothing," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, pp. 62–65, IEEE, New York, NY, 1981.

[42] H. Ney, "Dynamic programming algorithm for optimal estimation of speech parameter contours," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-13, pp. 208–214, March/April 1983.

[43] R. Bellman, *Dynamic Programming.* Princeton, N.J.: Princeton University Press, 1957.

[44] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans on Information Theory*, vol. IT-13, pp. 260–269, 1967.

[45] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, February 1975.