MULTITAPER ANALYSIS OF FUNDAMENTAL FREQUENCY VARIATIONS DURING VOICED FRICATIVES

Christine Shadle & Gordon Ramsay

School of Electronics & Computer Science, University of Southampton

ABSTRACT: A method for tracking fundamental frequency variations in speech is proposed, based on multitaper analysis. Using the multitaper technique, a statistical test is developed for detecting the presence of harmonic components at multiples of a fundamental frequency, embedded in coloured noise. It is shown that this can be applied to speech to estimate the fundamental frequency, when present, as well as the amplitude and phase of each harmonic. The method is validated on synthetic data, to determine accuracy and robustness, and evaluated on a small corpus of real speech data, comparing simultaneous acoustic and electroglottographic measurements to assess performance. Acoustic measurements are marginally less accurate than electroglottographic measurements, but often continue to provide useful fundamental frequency estimates in situations where electroglottography fails.

INTRODUCTION

Variations in the fundamental frequency of the voice (F0) are known to provide important perceptual cues in speech. Long-term fundamental frequency variations associated with different intonation patterns typically indicate higher-level linguistic or extra-linguistic information. However, short-term perturbations of the F0 contour also occur automatically when a constriction is formed in the vocal tract, or when vocal fold vibration is affected by aerodynamic factors; these perturbations are known to provide acoustic cues for individual consonants (Ohde, 1984; Baken & Orlikoff, 1988). Previous studies have used both acoustic and electroglottographic measurements to examine short-term fundamental frequency variations and devoicing phenomena in fricatives (e.g. Chollet & Kahane, 1979; Barry, 1995; Jesus & Shadle, 2003). Direct acoustic analysis of F0 during the production of consonants is often difficult to carry out reliably using traditional spectral estimation techniques. Existing methods are typically limited by tradeoffs in time and frequency resolution, and cannot track harmonics robustly in the presence of noise, which is characteristic of many fricatives and stops where F0 cues are important. Furthermore, mechanical vibration of the vocal folds may not always show up in the acoustic signal during obstruents. In cases where the vocal folds are in contact, electroglottography provides a more reliable indicator of voicing. However, it cannot always replace acoustic analysis for obstruents, since the vocal folds are partially abducted during these sounds, mechanical contact is reduced, and the EGG signal may be too weak to follow vocal fold vibration. Both acoustic and electroglottographic measurements are therefore needed, as well as improved signal processing methods for extracting voicing information from these data.

A new spectral analysis technique, multitaper analysis, has been developed recently, which significantly out-performs classical spectral estimators (Perceval & Walden, 1993). It relies on reducing spectral bias and variance through averaging of multiple eigenspectra, obtained by successively weighting any particular signal by a family of discrete prolate spheroidal sequences. A key result in multitaper analysis is the derivation of a rigorous statistical test for detecting harmonic signals in coloured noise (Thomson, 1982). The test involves evaluating the null hypothesis that only coloured noise is present at any frequency, against the alternative hypothesis that a complex harmonic masked by coloured noise is observed. By thresholding the test statistic at any desired significance level, harmonic components can be detected automatically across any frequency range of interest, even when non-white noise is present. The test is particularly appropriate for estimating fundamental frequency during sounds that are produced from a mixed excitation, such as voiced fricatives occurring in speech, where the time-varying harmonic structure that carries the pitch cues is buried in a coloured noise spectrum.

The aim of this paper, therefore, is to describe the development of a new technique for pitch tracking, based on multitaper analysis, which can be used to follow the fundamental frequency contour during fricative consonants. A simple extension of the basic multitaper harmonic analysis technique is derived, which tests for the presence of a sequence of harmonics at multiples of a single fundamental frequency. The proposed method is validated on synthetic signals with known F0 contours, embedded in increasing amounts of noise, in order to provide an estimate of the frequency resolution and assess robustness. Estimates of the F0 contour are then extracted independently from acoustic and electroglottographic signals recorded for a corpus of English phrases containing voiced and voiceless fricatives in vocalic contexts, to evaluate performance on real data.

MULTITAPER HARMONIC ANALYSIS

Let $X = \{X_n : n = 0, \ldots, N-1\}$ be a real-valued random process representing N samples of a speech signal, uniformly sampled in time. If the signal is assumed to be stationary, the spectral representation theorem indicates that it can always be decomposed into two components: a singular process with a line spectrum, made up of a series of complex harmonics, and a regular process with a continuous spectrum, made up of coloured noise. In analysing speech, the singular process represents the voiced component produced by the glottal source, whereas the regular process represents the unvoiced component, produced by supraglottal turbulent noise sources. The problem is to separate these two components, and to estimate the frequency, amplitude, and phase of each harmonic, and the spectrum of the coloured noise.

A survey of existing approaches for analysing voiced and unvoiced components in speech is given in Jackson & Shadle (2001). Traditional methods for source separation and fundamental frequency tracking often rely on directly examining the power spectral density of the process, which can be estimated using the Discrete Fourier Transform (DFT). However, the raw DFT typically suffers from poor bias and variance properties that can obscure spectral details in sounds such as fricatives. Bias can be reduced by pre-multiplying the signal by a data window. Variance can be reduced by smoothing the power spectral density estimates across frequencies or, equivalently, by multiplying the autocorrelation function by a lag window. Using a single data window and lag window, there is known to be an implicit tradeoff between bias and variance; reducing one typically increases the other. Better control of bias and variance can be achieved by combining power spectral density estimates obtained using several different data windows. This is the motivation for the multitaper spectral estimate introduced by Thomson (1982).

To define the multitaper estimator, let $\{h_{k,n} : k = 1, \ldots, K; n = 0, \ldots, N-1\}$ be a set of K window functions for X. For each window function, let $H_k : [0, 2\pi) \to \mathbb{C}$ be the spectral kernel of the window, calculated by taking the DFT of h_k:

$$H_k(\omega) \;:=\; \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} h_{k,n} e^{-j\omega n}, \tag{1}$$

and let $S_k^X : [0, 2\pi) \to \mathbb{C}$ be the direct spectral estimator of the signal, formed by taking the DFT of the product of h_k and X:

$$S_k^X(\omega) \;:=\; \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} h_{k,n} X_n e^{-j\omega n}. \tag{2}$$

The multitaper estimator $S_{MT}^X : [0, 2\pi) \to \mathbb{R}$ is defined to be the average of the K individual power spectral density estimates:

$$S_{MT}^X(\omega) \;=\; \tfrac{1}{K} \textstyle\sum_{k=1}^{K} |S_k^X(\omega)|^2. \tag{3}$$

If the set of window functions can be chosen so that (a) the spectral kernels $H_k(\omega)$ are maximally concentrated within a chosen bandwidth β about the origin, and (b) the direct spectral estimates $S_k^X(\omega)$ are uncorrelated at each frequency, then the multitaper estimator can be shown to reduce spectral bias and variance significantly with respect to each of the individual direct spectral estimators. A suitable choice is the family of discrete prolate spheroidal sequences investigated by Slepian & Pollak (1961); examples are illustrated in Figure 1, and the equivalent spectral kernel is shown in Figure 2.

Blacklock & Shadle (2003) have compared bias and variance properties of traditional and multitaper spectral estimates for fricative consonants. Both properties (a) and (b) can be exploited further to derive a statistical test for the presence of multiple harmonics in coloured noise, which is the subject of this paper. To do this, assume as above that the speech signal can be decomposed into voiced and unvoiced components. Suppose that the voiced component can be modelled as a sum of M harmonics located at integer multiples of a fundamental frequency ω_0, with amplitudes and phases defined by a set of complex coefficients $\{C_m \in \mathbb{C} : m = 1, \ldots, M\}$. Suppose that the unvoiced component can be modelled as a real-valued Gaussian coloured noise process $W = \{W_n : n = 0, \ldots, N-1\}$ with zero mean and continuous spectral density. Then:

$$X_n \;=\; \textstyle\sum_{m=1}^{M} (C_m\, e^{jm\omega_0 n} + C_m^*\, e^{-jm\omega_0 n}) + W_n. \tag{4}$$

Substituting (4) into (2) yields:

$$S_k^X(\omega) \;=\; \textstyle\sum_{m=1}^{M} (C_m\, H_k(\omega - m\omega_0) + C_m\, H_k(\omega + m\omega_0)) + S_k^W(\omega). \tag{5}$$

If the spectral bandwidth of the set of window functions is chosen to be less than the fundamental frequency, then for $\omega = m\omega_0$:

$$S_k^X(m\omega_0) \;\approx\; C_m\, H_k(0) + S_k^W(m\omega_0). \tag{6}$$

The resulting set of equations for $m = 1, \ldots, M$ and $k = 1, \ldots, K$ can be written in matrix form as:

$$\begin{bmatrix} S_1^X(\omega_0) & \cdots & S_1^X(M\omega_0) \\ \vdots & \ddots & \vdots \\ S_K^X(\omega_0) & \cdots & S_K^X(M\omega_0) \end{bmatrix} = \begin{bmatrix} H_1(0) \\ \vdots \\ H_K(0) \end{bmatrix} \begin{bmatrix} C_1 & \cdots & C_M \end{bmatrix} + \begin{bmatrix} S_1^W(\omega_0) & \cdots & S_1^W(M\omega_0) \\ \vdots & \ddots & \vdots \\ S_K^W(\omega_0) & \cdots & S_K^W(M\omega_0) \end{bmatrix}, \tag{7}$$

which is equivalent to a standard complex linear regression model:

$$\mathbf{X} \;=\; \mathbf{HC} + \mathbf{W}. \tag{8}$$

Since the signal and noise processes are assumed to be Gaussian, the spectral estimates obtained using each window function will also be Gaussian. If the window functions are discrete prolate spheroidal sequences, then it can further be shown that the spectral estimates obtained using different window functions will be approximately uncorrelated. Under these assumptions, the maximum likelihood estimator $\hat{\mathbf{C}}$ for the complex vector of harmonic coefficients $\mathbf{C}$ is given by:

$$\hat{\mathbf{C}} \;=\; (\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*\mathbf{X}. \tag{9}$$

The statistics of the maximum-likelihood estimator and the corresponding error covariance matrix can easily be determined, and used to construct a statistical test for the presence or absence of the harmonic components. Following the argument outlined in Thomson (1982), it is not difficult to show that, if all of the M harmonic components are identically zero,

$$\sigma := \frac{(K-1)\sum_{m=1}^{M}\sum_{k=1}^{K}|H_k(0)\,\hat{C}_m|^2}{\sum_{m=1}^{M}\sum_{k=1}^{K}|S_k^X(m\omega_0)-H_k(0)\,\hat{C}_m|^2} \sim F_{2M,2M(K-1)}, \tag{10}$$

where $F_{2M,2M(K-1)}$ is the F-distribution with $2M$ and $2M(K-1)$ degrees of freedom. This defines a standard variance ratio test (F-test) with test statistic σ. The numerator of the expression for σ represents the total power in the harmonic components, whereas the denominator represents the total power in the residual coloured noise at the same frequencies, so the test essentially compares these two quantities. The null hypothesis that only coloured noise is present in the signal, with no harmonic components at multiples of the fundamental frequency ω_0, can be rejected at significance level α whenever $\sigma > F_{2M,2M(K-1)}(\alpha)$, leaving the alternative hypothesis that both coloured noise and at least one harmonic of the fundamental are present whenever $\sigma \leq F_{2M,2M(K-1)}(\alpha)$.

The basic test formulated above enables the null hypothesis to be accepted or rejected at a fixed significance level for a single predetermined fundamental frequency. In practice, the fundamental frequency is unknown, and must be located by constructing a series of tests over an appropriate grid of candidate values, sampled finely-enough at regular intervals along the frequency axis to guarantee the desired frequency resolution. If none of the resulting test statistics is found to be significant, then the signal is considered to be unvoiced. If one or more of the test statistics is significant, then the results of the test with the greatest significance level are taken to indicate the best fundamental frequency value and corresponding harmonic coefficients.

The test also assumes that the signal is stationary, but this is only true over short time-intervals for speech. Temporal variations in fundamental frequency are specifically of interest here, and need to be analysed. This can be done in the usual manner by segmenting the signal into overlapping frames of suitable duration, and applying the multitaper harmonic analysis technique to estimate the fundamental frequency and harmonic spectrum for each frame in turn. No continuity constraints on the fundamental frequency contour are necessary, although the F-test typically gives spurious results in the middle of voiceless consonants, when the signal level drops too far. Performance can be improved if a simple energy threshold is used to disable the test as soon as voicing can no longer be detected, until the signal energy rises above an appropriate threshold again.

METHOD

Two preliminary studies were conducted to assess performance. The first study was designed to validate the behaviour of the multitaper analysis procedure using a corpus of synthetic test signals. Each test signal was composed of a single frequency-modulated sinewave of constant amplitude, embedded in Gaussian white noise of variable amplitude. To test the sensitivity of the procedure to rapid variations in fundamental frequency, the instantaneous frequency of the sinewave was made to vary sinusoidally in time about a centre frequency c_{cf}, with modulation frequency c_{mf} and modulation amplitude c_{ma}. To test the robustness of the procedure against additive noise, the signal-to-noise ratio c_{snr} was made to vary by altering the noise amplitude. The default parameter values were c_{cf}=100 Hz, c_{mf}=5 Hz, c_{ma}=20 Hz, c_{snr}=∞. To generate the entire corpus, each parameter was varied independently in six steps from its default value, keeping the other parameters constant, to give a total of 24 signals, all of which were calculated over a duration of 1 second using a 16 kHz sampling rate. The multitaper analysis procedure was then applied to each signal in the corpus to determine fundamental frequency estimates with 0.1 Hz resolution at 1 ms intervals over 35 ms frame lengths, using 3 prolate spheroidal data tapers with an effective bandwidth of 65 Hz. The results were compared with the original instantaneous frequency trajectories used to generate the test signals, resampled at corresponding 1 ms intervals. The percentage of samples exhibiting gross errors, where the estimated fundamental frequency deviated by more than 20% from the true value, was determined. The r.m.s. discrepancy between true and estimated values was calculated for the remaining samples.

The second study was designed to evaluate the performance of the multitaper analysis procedure on real speech data, comparing results obtained using both acoustic and electroglottographic measurements. A single male speaker of British English was made to produce utterances consisting of $/CV_1CV_2C/$ nonsense words, with the stress on the first syllable, embedded in the carrier phrase “Say a __ again”, for all combinations of fricative consonants $C \in$ {f, θ, s, ʃ, v, ð, z, ʒ} and vowels $V_1, V_2 \in$ {ae, i, u}. Simultaneous acoustic and electroglottographic recordings were made for ten repetitions of each utterance. The acoustic signal was recorded using an AKG C419 condenser microphone and sampled at 16 kHz. The electroglottographic signal was recorded using a laryngograph produced by Laryngograph Ltd (UK), and sampled at 2 kHz, then resampled to 16 kHz. Both signals were quantized using linear PCM encoding with 12-bit resolution. To provide a reliable reference estimate of the fundamental frequency contour, the EGG signal was bandpass-filtered between 50 Hz and 750 Hz, and an automatic peak-picking algorithm was used to detect glottal closure events by locating maxima in the EGG signal and its derivative. Fundamental frequency estimates were calculated as the reciprocal of the time between two adjacent glottal events during each voiced interval, and positioned mid-way between each pair of events. Spline interpolation was then used to resample the resulting fundamental frequency contour uniformly at 1 ms intervals over the intervals in which voicing could be detected, setting values to zero elsewhere. The multitaper analysis procedure was subsequently applied to both the acoustic and electroglottographic signals to determine fundamental frequency estimates with 0.1 Hz resolution at 1 ms intervals over 35 ms frame lengths for each utterance in the corpus, using 3 prolate spheroidal data tapers with an effective bandwidth of 65 Hz. The fundamental frequency contours obtained from each type of signal were then compared individually with the reference contour calculated by detecting events in the EGG signal. The percentage of samples exhibiting gross errors, where the frame-based fundamental frequency estimate deviated by more than 20% from the event-based estimate, was determined. The r.m.s. discrepancy between frame-based and event-based estimates was calculated for the remaining samples, and expressed as a percentage of the mean value. Also calculated were the percentage of samples where the frame-based procedure failed to provide estimates that were provided by the event-based procedure (omissions), and conversely where the frame-based procedure provided estimates that were not provided by the event-based procedure (additions).

RESULTS

The results of the first study are shown in Figures 3-5. The accuracy of the multitaper analysis procedure approaches the chosen fundamental frequency resolution of 0.1 Hz for the default parameter values. The error rate increases slightly as the fundamental frequency of the signal increases, and rises noticeably as the fundamental frequency varies more rapidly relative to the frame length; this is typical of frame-based estimation procedures, and limits the detection of rapid pitch changes. The estimates are robust against additive noise; no gross errors are seen until the signal-to-noise ratio approaches -10 dB, when the r.m.s. discrepancy is around 4% . All of these results confirm that the multitaper analysis procedure performs well, and is capable of accurately tracking fundamental frequency variations in noise.

Results from the second study are given in Figure 6. The total r.m.s. discrepancy for the frame-based estimates obtained from the EGG signals was 1.77 Hz (1.83%), compared to 2.38 Hz (2.43%) for the acoustic signals. Very few gross errors were observed for either signal type; only 0.0224% of samples for the EGG signals, and 0.0288% for the acoustic signals. Estimates calculated using the laryngograph signal are thus systematically better than estimates obtained using the microphone signal, as expected, but both methods provide reliable results. Frame-based methods typically underestimate rapid transitions near voicing boundaries, and results obtained from both signal types were worse for voiceless than for voiced fricatives. Comparing event-based fundamental frequency estimates using the EGG signal with frame-based estimates obtained from the acoustic signal, the multitaper analysis procedure was found to omit 9.42% of the samples that were obtained by glottal event detection, but added 7.89% that were missed. This demonstrates that multitaper analysis of microphone recordings may provide a useful complement to laryngograph measurements for estimating fundamental frequency contours in sounds such as fricatives.

As an illustratation of the application of multitaper analysis to continuous speech, Figure 7 shows a multitaper power spectrogram of the phrase “Weatherproof galoshes are very useful in Seattle”, taken from the TIMIT corpus, with frames calculated at 1 ms intervals over 10 ms frame lengths; 4 data tapers were used to construct each estimate, with an effective bandwidth of 250 Hz. Figure 8 shows a multitaper F-test spectrogram for the same phrase, with frames calculated at 1 ms intervals over 35 ms frame lengths; 4 data tapers were used to determine each statistic, with an effective bandwidth of 75 Hz, and tests were constructed for a single fundamental frequency component over the range 0-2 kHz at intervals of 1 Hz. The multitaper power spectrogram is comparable to a traditional wide-band spectrogram, except that the reduction in variance afforded by the multitaper procedure results in a smoother image quality. Typically, spectral details are sharpened but also broadened by the rectangular spectral kernel, so formant frequencies are clearer but formant bandwidths may be artificially enlarged. Thus, multitaper spectral estimates are not usually suitable for locating precise peaks, such as formants or harmonics. On the other hand, the multitaper F-test spectrogram resembles a traditional narrow-band spectrogram, but the harmonics are much sharper and the unvoiced portions of speech do not appear; consequently, the F-test statistic is usually better for locating harmonics than a direct spectral estimate. Figure 9 shows the F0 contour and the amplitudes of the first five harmonics estimated using the multitaper harmonic analysis procedure for the same phrase; tests were constructed at the 10% significance level for 10 harmonic components of a fundamental frequency lying in the range 0-200 Hz, at intervals of 0.5 Hz. Individual harmonics are tracked throughout each consonant, and localized perturbations in frequency and amplitude are clearly visible during periods of oral constriction. The envelope of the harmonic spectrum is determined by the transfer function of the vocal tract and the glottal source spectrum; higher harmonics often appear to cut off before lower harmonics at fricative boundaries, which may indicate a change in the spectral slope of the source as vocal fold vibration is inhibited. Information about individual harmonics is not available when using traditional pitch-tracking algorithms, but can be obtained automatically using multitaper analysis.

CONCLUSIONS

A novel technique has been developed for estimating the fundamental frequency contour in continuous speech, using multitaper analysis. Unlike traditional pitch-tracking algorithms, the time-varying amplitudes and phases of individual harmonics can be obtained using the proposed procedure, and used to analyse voicing. The method has been validated on synthetic data, and shown to be both accurate and robust. Results from a small corpus of continuous speech data indicate that fundamental frequency contours extracted from the acoustic signal are comparable to fundamental frequency contours estimated from electroglottographic waveforms; acoustic estimates are typically $\sim$ 1% less accurate than EGG-derived estimates, but EGG-derived estimates may fail during consonants when glottal abduction occurs. Multitaper harmonic analysis therefore provides a useful complement to electroglottography. Future work will extend the technique further to provide a comprehensive methodology for automatically separating and characterizing voiced and unvoiced components of the speech signal.

REFERENCES

Baken, R. J. & Orlikoff, R. F. (1988) “Changes in vocal fundamental frequency at the segmental level: control during voiced fricatives” *Journal of Speech and Hearing Research* **31**(2), 207–211.

Barry, S. M. E. (1995) “Variation in vocal fold vibration during voiced obstruents in Russian” *European Journal of Disorders of Communication* **30**, 124–131.

Blacklock, O. S. & Shadle, C. H. (2003) “Spectral moments and alternative methods of characterizing fricatives” *Journal of the Acoustical Society of America* **113**(4/2), 2199.

Chollet, G. F. & Kahane, J. C. (1979) “Laryngeal patterns of consonant productions in sentences observed with an impedance glottograph” in H. Hollien & P. Hollien (eds) *Current Issues in the Phonetic Sciences*, John Benjamins, Amsterdam, 119–128.

Jackson, P. J. B. & Shadle, C. H. (2001) “Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech” *IEEE Transactions on Speech and Audio Processing* **9**(7), 713–726.

Jesus, L. M. T. & Shadle, C. H. (2003) “Devoicing measures of European Portuguese fricatives” in N. J. Mamede, J. Baptista, I. Trancoso & M. das Graças Volpe Nunes (eds) *Computational Processing of the Portuguese Language*, Springer Verlag, Berlin, 1–8.

Ohde, R. (1984) “Fundamental frequency as an acoustic correlate of stop consonant voicing” *Journal of the Acoustical Society of America* **75**(1), 224–230.

Perceval, D. B. & Walden, A. T. (1993) *Spectral Analysis for Physical Applications*, Cambridge University Press, Cambridge, U.K.

Slepian, D. & Pollak, H. O. (1961) “Prolate spheroidal wave functions, Fourier analysis and uncertainty - I” *Bell Systems Technical Journal* **40**, 43–63.

Thomson, D. J. (1982) “Spectral estimation and harmonic analysis” *Proceedings of the IEEE* **70**, 1055–1096.

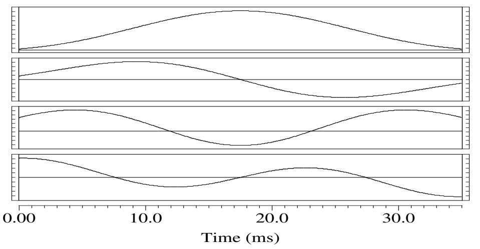

Figure 1: Window functions for DPSS tapers, BW=100 Hz.

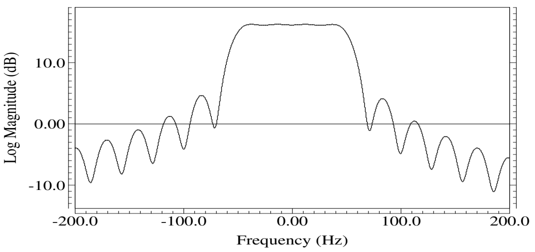

Figure 2: Spectral kernel for DPSS tapers, BW=100 Hz.

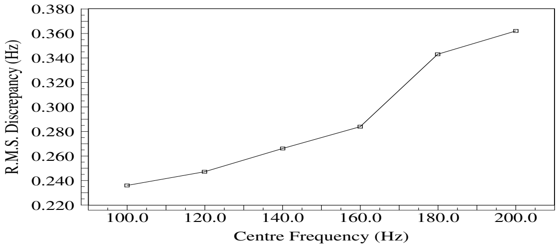

Figure 3: Effect of centre frequency.

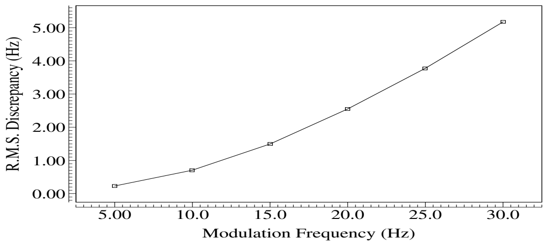

Figure 4: Effect of modulation frequency.

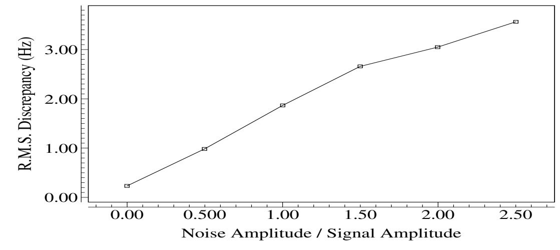

Figure 5: Effect of noise-to-signal ratio.

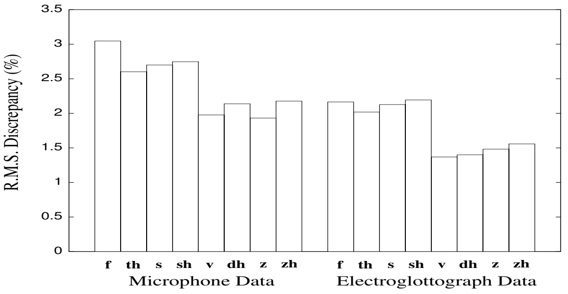

Figure 6: Error rates for EGG and microphone data.

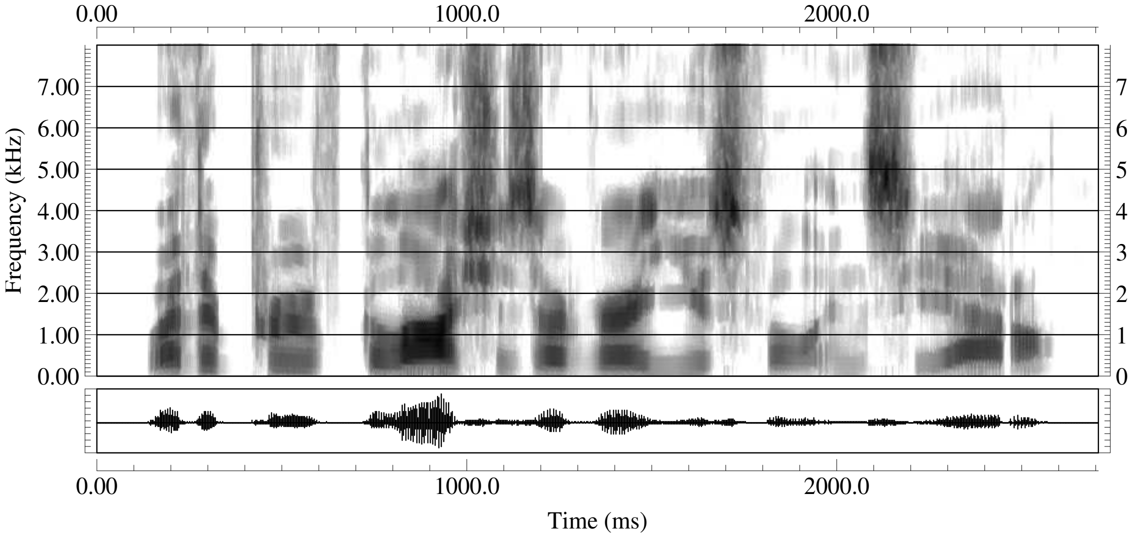

Figure 7: Multitaper spectrogram constructed using the multitaper power spectral density estimate.

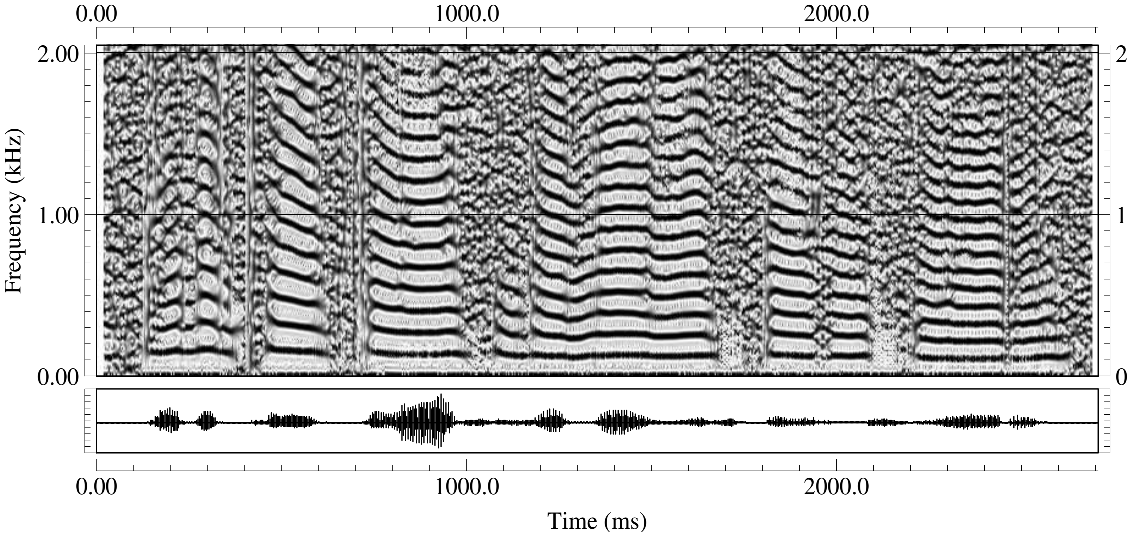

Figure 8: Multitaper spectrogram constructed using the multitaper F-test statistic.

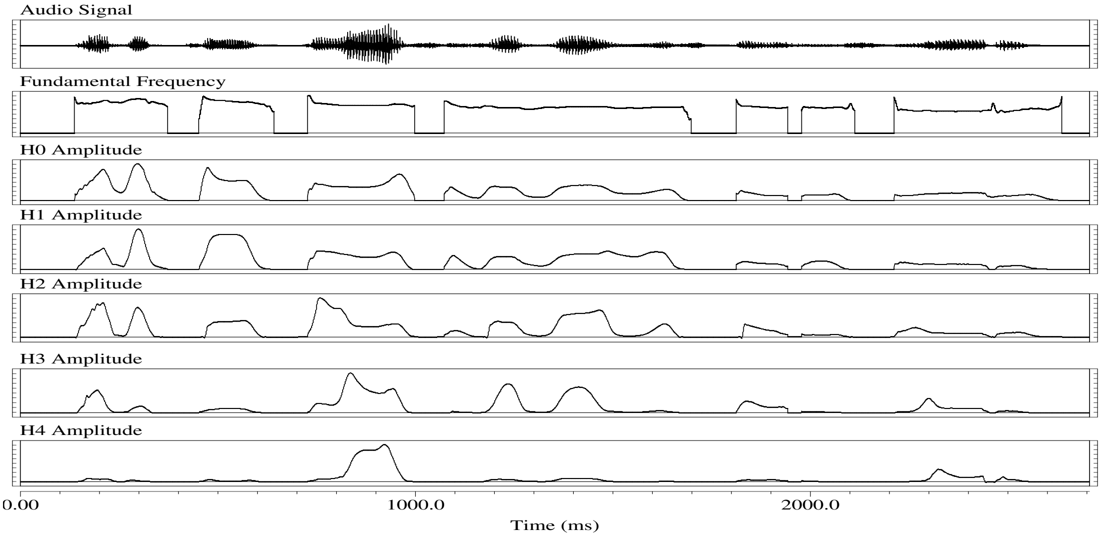

Figure 9: Multitaper harmonic analysis showing estimated F0 contour and harmonic amplitudes.