A Quasi-glottogram signal for voicing and power estimation.

Greg Kochanski and Chilin Shih

Bell Laboratories, Lucent Technologies

Abstract

We propose a novel, noninvasive experiment that reliably shows the strength of glottal oscillations. The Quasi-glottogram (QGG) signal is generated from a microphone array and an electroglottogram signal. It can be used to improve estimates of whether speech is voiced, quantify partial voicing, and reduce the phoneme effect in speech signals. The technique is well adapted to the generation of text-to-speech systems, as it allows an estimate of the glottal flow during undisturbed, natural speech. For prosody studies, it can be used to provide an estimate of amplitude which is relatively unaffected by changes in phonemes, at least as reliable as standard estimators of amplitude.

PACS: 43.70.Jt, 43.70.Fq, 43.70.Aj, 43.58.Ta, 43.38.Hz

Abbreviated title: “A Quasi-glottogram signal”

Corresponding author: Greg Kochanski gpk@bell-labs.com

1. Introduction

The source-filter model of speech production explains the acoustic speech signal as a convolution of the time-varying glottal airflow (due to the vibrations of the vocal folds) with the impulse response of the acoustic filter formed by the vocal tract. A measurement of the glottal airflow is clearly desirable: it would allow independent studies of the source and the filter. In particular, it would allow a determination of voicing, allowing measurements of the amplitude and harmonic content of the glottal oscillations. The source-filter model of speech is embedded in speech coders, automatic speech recognition systems and speech synthesizers; better understanding of the source and filter separately can lead to better algorithms. For example, speech synthesizers need reliable, precise indications of glottal oscillation because voiced speech may be processed differently from unvoiced speech. Human listeners will easily detect errors in the database of a synthesizer where the speech has a voicing indication that is wrong for more than 30 ms, a sensitivity that requires improvement in current techniques if one wants to create synthesizers automatically without tedious manual checking.

Further, a relatively unexplored area of speech science is the study of prosody. Prosody includes all the acoustic properties of speech beyond the sequence of phonemes, such as the pitch, the amplitude, and the spectral tilt of the sound. Prosody is used to mark boundaries, to emphasize words or phrases, to help control dialogs, among other functions. However, quantitative measurements of prosody are non-trivial, because other than pitch, all the candidate acoustic features are strongly influenced by the phonemes: an emphatic ‘m’ may be quieter than a soft ‘a’. Thus, to be able to study and understand prosody, there is a need for measurements of prosody that are relatively independent of the particular phonemes, so that one can compare prosody of different words. Again, measurements of the glottal flow would be useful, because on one hand, the glottal flow is much less dependent on the choice of phoneme than the far-field acoustic signal is, and on the other hand because it can be related to physiological parameters like subglottal pressure and muscle tensions.

Invasive estimates of the glottal flow are possible, using acoustically matched tubes [1,2], but this interferes with lip and jaw movements. This interference is fatal to applications which require simultaneous recording of natural, undisturbed speech. One such application is the recording for a text-to-speech-system (TTS) database. Most commercial TTS systems are built by piecing together segments of real speech. Natural, high quality recordings are a basic requirement: but one also needs to automatically and reliably estimate acoustic properties of the speech signals. Other estimates of glottal parameters, such as intubation to measure subglottal pressure [3], direct photography of the vocal folds [4,5], or photoglottography [6,7] to measure the glottal open area are also invasive and incompatible with sensitive applications. Plethysmography [8] can measure subglottal pressure, but is cumbersome, and the acoustic properties of the box in which the subject sits need to be carefully considered in order to get clean speech. Finally, Electroglottography (EGG) [9,10,11,12,13,14,15,16] is noninvasive, but produces only an indication of vocal fold contact, and the folds can be contacting while the glottis is partly open, or non-contacting when the glottis is nearly closed. The EGG signal doesn’t usefully measure the width of the glottal opening, and thus misses most of the information in the source waveform.

Historically, the glottal flow has been estimated by “inverse filtering.” Inverse filtering is based on first estimating the vocal tract transfer function from a microphone or anemometer signal, inverting it, and using the inverted filter to remove the effect of the vocal tract [17]. If the estimated transfer function is accurate, the result will be close to the actual glottal flow. However, the problem is intrinsically indeterminate, as one is trying to estimate both a time series of glottal flow and a time series of vocal tract parameters from a single input time series. In practice, assumptions are made that the microphone signal is quasistationary, that the spectrum of the glottal flow is simple near the vocal tract’s formant frequencies, that the vocal tract can be modeled as an all-pole spectrum, and sometimes that the formant frequencies change smoothly with time. While the resulting algorithms work reasonably well, none of the approximations are perfect, and the resulting estimate of the glottal flow is only approximate, with ill-determined errors.

The goal of this study is to develop a new, noninvasive technique that allows an estimate of the volume flow of air through the glottis, U, based on an array of microphones and an EGG signal. This paper will first justify why such an algorithm is possible. Second, it will lay out the details of our algorithm. Third, we test the algorithm. Since we do not have invasive flow measurements available for a direct comparison, we bring in two lines of indirect evidence to show that the Quasiglottogram (QGG) signal is closer to *U* than the standard signals used for voicing estimation. The first test is qualitative: we compare the QGG’s behavior to other signals in “difficult” regions of speech, and show that the QGG behaves well under conditions where one or another of the standard signals misbehaves. The second test is quantitative, though indirect: To prepare, we introduce a simple “toy” model of speech, and show that in that model, the signals that lead to the least variable amplitude estimates are the signals that are closest to *U*. We then show that the QGG signal allows a very steady estimation of amplitude, generally less variable than the result of other standard linear estimators. This provides evidence that the QGG signal is close to *U*.

A good estimator for glottal flow should be noninvasive, linear, and should be related to the actual glottal airflow through a time-invariant transfer function. It should also distinguish between voiced sounds and sounds generated in other constrictions of the vocal tract. Linearity helps simplify connecting the estimate of *U* to physiologically important parameters like the subglottal pressure and the glottal open area and it allows straightforward quantification of partial voicing. The output of the algorithm should be related to *U* via a time-invariant transfer function, so that we can meaningfully compare signals at one time to signals at another. Particularly, one would like to compare glottal flows between different phones.

We approach these goals by building a signal from linear combinations of several filtered microphone signals. We choose the linear combinations and filters to make the best possible match to the EGG signal, which provides an instantaneous measurement of whether the vocal folds are contacting or not, and thus gives some indication whether the glottis is open or closed. We also use data from a microphone near the base of the throat to pick up a signal from the subglottal cavities. The subglottal cavities have an acoustic transfer function from the glottis through the throat wall to an outside microphone that is relatively independent of time, as (unlike the vocal tract) the trachea is not surrounded by muscles used to articulate speech. The dimensions of the trachea and bronchi are also largely unchanged during breathing, and even motions of the larynx are expected to lead to only modest changes in the acoustic resonances [18,19] of the trachea.

The algorithm requires at least two microphones. One to pick up the throat signal, and one for sound from the vocal tract. The throat signal contains a mixture of sound that propagated from the glottis down through the trachea then through the skin of the throat, mixed with sound that propagated from the glottis up through the vocal tract, then back down through free air to the throat microphone. The other microphones are used primarily to cancel out the component from the vocal tract. This cancellation results in a signal that is the glottal waveform, filtered by its propagation through the neck. We show that this signal is less variable than commercial inverse-filtered signals, in the sense that the signal shows less phoneme-dependent variation. This stability is what one expects of the glottal waveform; it should be only weakly influenced by the vocal tract configuration [20].

It is necessary to cancel out the sound from the vocal tract because we want an estimator of the glottal flow, and we don’t want our measurement to be disturbed by the dramatic changes in the transfer function of the vocal tract that occur during normal speech. This is a different approach from an inverse-filter estimator, which attempts to dynamically estimate and invert the vocal tract transfer function: a nontrivial, multiplicative operation on one signal. We look for a signal which has a time-invariant relationship to the glottal flow via a linear operation on several signals.

2. Experimental Apparatus

In the experiments described below, we mounted four Bruel and Kjaer type 4165 omnidirectional microphones on the face guard of a hockey helmet. The microphones were mounted near the nose (4 cm lateral from the end of the nose), to the side of the mouth (2 cm lateral of the corner of the mouth), near the forehead (centered, 11 cm above nose level), and near the base of the throat (on centerline, 2 cm from skin, 2 cm above the top of the sternum). The microphones were checked to be slightly outside breath streams, and were protected by 4 mm of windscreen foam. In other experiments, we found that only the placement of the throat microphone was critical: it should be placed as close as possible to where the trachea can be palpated (the *fossa jugularis*), so long as the microphone does not collide with the subject during normal head motions. The placement of the other microphones is not critical, and we have obtained similar signals from a 6-microphone array (including cheek and back-of-the-head microphones), and a 3-microphone array (using a gradient microphone for the throat signal).

The EGG signals were obtained from a Portable Laryngograph [21]. The electrodes were placed to maximize the signal strength for long vowels, in modal speech, over the normal range of f_0. Data was digitized at 12 kHz per channel with an antialiasing filter. The microphone and EGG signals are high-pass filtered at 40 Hz to reduce room noise and the large amplitude, slow components of the EGG that correspond to motions of the larynx. All inverse filtered signals were produced by ESPS/waves [22].

In the examples that follow, we used Mandarin speech acquired from a female subject, a native Mandarin speaker fluent in English. Because Mandarin is a tone language, Mandarin speakers can conveniently control f_0 movements. Therefore we expect that vocal folds vibration will be more repeatable, reducing unpredictable fluctuations in *U*. The subject pushed a key for each prompt, and Chinese characters were presented on a screen. Speech began 0.72±0.1 seconds after the prompt.

The data presented in section 4 was taken from a random (*e.g.*, selected for other purposes) set of English vowels and words, read at normal volume in modal speech, along with some unplanned spontaneous phrases. The data for section 5 is a database of 979 utterances in the form *"Ta1 shuo1 X san1 tian1."* ("He says *X* for three days."), where *X* ranges over a random selection of syllables in all four tones. Overall, f_0 had a mean of 235 Hz and a standard deviation of 26 Hz, with most of the variation on *X*.

Calibration utterances to train the algorithm (*i.e.*, fit the filter coefficients), for both sections, were a broad mixture of sounds. We recorded 25 calibration utterances at the beginning of the day's recording, and another 25 at the end, 6 hours later. Half were single-syllable English words (*e.g.*, "nap"). One quarter were repeated fricatives, 'p', and 'h'. The remainder were long vowels, nasals, and voiced fricatives where f_0 was swept through the subject's comfortable pitch range. Voiceless sounds are useful to force cancellation of the mouth signal and to check the operation of the algorithm, because their *U* is nearly a steady flow. The f_0 sweep sentences provide voiced speech with the speaker's full pitch range. It is useful when the calibration utterances cover all the f_0 range of the speech used in the experiments. The calibration utterances did, having range of f_0 with mean 248 Hz and standard deviation 97 Hz.

The FIR filters used in the QGG algorithm had taps spaced over 13 milliseconds.

3. The algorithm

The overall structure of the algorithm is shown in Figure 1. To explain the design of the algorithm, we can consider a simplified form of the microphone array as shown in Figure 2: one microphone, M, near the mouth, and another, T, at the base of the throat. During voiced sounds, the signals are excited by the flow, *U*, through the glottis. The signal at T is made of two main components, one traveling directly through the neck [23] (*via* a transfer function *N*), and the other through the vocal tract (transfer function *V*) to the vicinity of M, then down through the air (transfer function *A*) to T. (Here, we treat these transfer functions as generic matrix manipulations in any complete basis, not yet specializing to a time series, frequency representation, or some intermediate basis, following [24].) In this model, $T = (N + A \cdot V) \cdot U$. We take $M = V \cdot U$, neglecting the small amount of sound coming through the neck.

We expect that *N* and *A* should be nearly independent of the phoneme being produced. On the other hand, V varies dramatically and systematically as a function of the phoneme. We can express this variation by writing $V = \bar{v} + \alpha \cdot \tilde{v}$, where V depends on the phoneme only through α , and $\langle \alpha \rangle = 0$ (the angle brackets denote an average over the corpus of speech). Because human speech is the result of several independently controlled articulators, α is normally a vector. In other words, we decompose V into an average transfer function, $\bar{v}$, and the variations around the average. Then, $T = (N + A \cdot \bar{v}) \cdot U + \alpha \cdot A \cdot \tilde{v} \cdot U$.

Hypothetically, if we had a reference signal, Q, which was derived from U via some time-invariant transfer function, q, we could find a linear filter that would reproduce Q from the microphone signals (note that we assume off-line processing, so our filters need not be causal). The general form for that linear filter would be $C = c_M \cdot M + c_T \cdot T$, and to find the filter matrices (coefficients) c_M and c_T, we would minimize the difference between C and Q: $c_M, c_T = \arg\min E(c_M, c_T)$, where

Eq. 1 $$E(c_M, c_T) = (C-Q)^T \cdot (C-Q) = \left\| C - Q \right\|^2$$

is the sum of squared errors between our target signal Q and our reconstruction.

If $\alpha = 0$, Eq. 1 is degenerate, and its solutions are all filter coefficients c_M and c_T on a line which we will call **L**, that goes through $c_{M1} = 0$, $c_{T1} = q \cdot (N + A \cdot \bar{v})^{-1}$ and $c_{M2} = q \cdot \bar{v}^{-1}$, $c_{T2} = 0$. The first of these two solutions corresponds to using just one filter, on the throat microphone to match *Q*, while ignoring *M*. The second solution is the reverse: using just one filter on the mouth microphone while ignoring *N*. Points on **L** correspond to linear combinations of these two filters, and all points on **L** are equally good solutions and match $C = Q$ exactly, assuming that the necessary inverses exist. This derivation can be extended to allow white additive noise in the microphones, in which case the necessary inverses always exist. Results are qualitiatively similar, although the derivation and results become substantially more complex.

If the speaker starts talking, instead of just vocalizing with a stable vocal tract, the transfer functions will vary from phoneme to phoneme, and we will not be able to match *Q* perfectly at every moment, so *E* will be positive. Not every solution gets the same increment of error, though. Solutions that depend predominantly on *V* will fit worse and have larger errors than solutions that depend predominantly on *N*, because *V* changes from phone to phone. This difference breaks the degeneracy and typically picks out a single best solution. In general, the best solution will use signals from all microphones, and it will provide a better approximation to *Q* than could be obtained from any linear filter operating on any single microphone in the array.

If we assume that *V* varies slowly enough, we can write down the change in error due to the difference between *V* at a given moment and the average of *V* (*i.e.,* $\bar{v}$):

Eq. 2 $$\Delta E = \left\langle \left\| (c_T \cdot A + c_M) \cdot \alpha \cdot \tilde{v} \cdot U \right\|^2 \right\rangle,$$

where the average (written as angle brackets) is taken over all phonemes in the corpus. The change in error is always non-negative and is normally nonzero everywhere except on a hyperplane we will call **P**, defined by $c_M = -c_T \cdot A$. Not coincidentally, this relationship between the filter coefficients is exactly what is needed to cancel out the part of the signal that came from the mouth, leaving only the part that came through the throat wall. The intersection of **L** and **P** then specifies the best estimator for *Q* (the one that is least sensitive to changes in the transfer functions). These results generalize to arrays of more than two microphones. They also can be generalized to the case where all the transfer functions vary with time, though Eq. 1 and Eq. 2 will change in detail.

So, given a reference signal, we can build an estimator of the glottal airflow by using an array of microphones, and finding the linear combination of filtered microphone signals that best matches the reference signal. The resulting signal is less variable than any signal derived by a linear operator from a single microphone. Loosely speaking, the best estimator is obtained by canceling out the highly variable signal from the mouth microphone, and using the part of the signal that did not propagate through the vocal tract.

In the real world, one doesn't normally have a perfect reference signal, Q. The best we can obtain non-invasively is the EGG signal. The EGG is related to U in a nonlinear and variable manner, because the larynx moves up and down relative to the electrodes used to measure the EGG. Repeating the above analysis shows that the variability of EGG measurements are not important, so long as changes in the EGG signal are not correlated with phonemes. This is true, by and large, as the larynx moves in response to pitch changes and inhalations, neither of which are correlated with most phonemes in most languages. Glottal stops and pharyngeal sounds are an exception, but they are not particularly common, typically comprising just a few phonemes in a language.

Nonlinearities in the relationship between U and the EGG signal are difficult to analyze analytically. We have investigated their effect empirically, in Sections 4 and 5.

3.1. Algorithm Introduction

We first build the data matrix, X, where each row contains the signal from one of the microphones, and each column corresponds to one moment in time. X is an n by m matrix when there are n microphones, each digitized to produce m samples of audio. The EGG signal is a one by m matrix.

We then select a set of taps (*i.e.*, taps on a delay line) for each microphone. Physically, the closure of the glottis is the cause of the acoustic signals: when it closes, a sound wave propagates up the vocal tract and down the trachea, reaching the microphones roughly a millisecond later. If one tries to observe a glottal closure at time t, one will need to use microphone data from later times, when the sound waves from the closure reach the microphones. Thus, we use the taps to build a finite impulse response (FIR) filter to predict the present EGG signal from future microphone signals.

To select the span of the taps, we need to consider the purposes of the filter we are building. It needs to cancel out the mouth signal that is picked up by the throat microphone, and it needs to match the impulse response of the remainder of the throat microphone signal to the EGG signal.

The taps span the range of delays beginning with the earliest propagation from the glottis to the microphone in question, ending when the impulse response of the vocal tract goes below 1% of its peak value.

Vocal tract formant bandwidths can be as small as 40 Hz [25,26] when the glottis is closed, though in real speech a bandwidth of 80Hz is more realistic [27]. Such a bandwidth implies that the vocal tract resonances will take about $3 \cdot 1/(2\pi B) \approx 6\text{ms}$ to decay. Bandwidths for the tracheal resonances are wider, 200-400Hz [28], and so are not the limiting factor for the window width in the time domain. Matching the acoustic to the EGG signal requires a window length of about $1/f_0$, which can be slightly longer. We choose the range of taps to cover the longer of these times, and we use the same number of taps for every microphone.

One needs enough data to capture a wide range of speech conditions, especially speech with a range of fundamental frequencies (f_0). The algorithm will not produce good estimates for speech that has f_0 outside the range of the training data in x.

The QGG signal does have some imperfections. It has a small response to fricatives and plosives. The primary source of that response is incomplete cancellation of the mouth acoustic signal at the throat microphone. The design of our microphone array allows the microphones to move about 1 mm relative to the skull. This small movement is expected to lead to 3% changes in value of the transfer functions, which would make cancellation of the mouth signal impossible to better than a 3% level. It also does not estimate U directly, but estimates U times a transfer function, where the transfer function depends on the subject and the experimental configuration.

Note that we do not claim any absolute calibration for the QGG signal. Because the QGG filter coefficients involve the EGG signal, the QGG amplitude will differ from person to person and session to session, depending on neck structure and the placement of the EGG electrodes. However, the QGG signal depends on the EGG only through its average properties during the calibration/training session. So, when one is actually using the QGG (as opposed to calibrating it), the EGG is entirely unused, and may be disconnected. Therefore, factors that affect the EGG signal during use (such as the momentary position of the larynx with respect to electrodes) will have no effect on the QGG.

3.2. Implementation

We build a set of linear equations corresponding to the FIR filter that best predicts the EGG signal, using straightforward least-squares linear prediction techniques[29]. The predicted EGG signal at each time is a linear combination of $n \cdot q$ values (q taps on each of n microphones). To start, we define the covariances between shifted signals from the i^{th} and j^{th} microphones:

Eq. 3 $$\phi_{\alpha,i,j} = \tfrac{1}{m}\sum_t x_{i,t-\alpha} \cdot x_{j,t} \text{ ,}$$

where α is the time shift between the i^{th} and j^{th} microphones (we neglect end effects, for simplicity), and t indexes the time. These covariances are estimated from the data, and contain noise covariances. Analogously, we will write $\phi_{*,\alpha,j}$ for the covariances between the EGG signal and the shifted microphone signals. The filter coefficients that minimize the mean squared error are then the solution to

Eq. 4 $$\sum_i \phi_{\alpha,i,j} \cdot c_{\alpha,i} = \phi_{*,\alpha,j} \text{,}$$

which is a set of $n \cdot q$ linear equations. We prepare to solve the equations by stacking the $c_{\alpha,i}$ to make a single vector *C*, stacking $\phi_{*,\alpha,j}$ to make a single vector *P* (of size $n \cdot q$ by 1), and placing the elements $\phi_{\alpha,i,j}$ into the corresponding places $H_{\alpha n+i,j}$ to make a $n \cdot q$ by n matrix. We solve the resulting matrix equation, $HC = P$, with a singular value decomposition algorithm to allow for degeneracies and near-degeneracies. *C* can then be unpacked to yield the $c_{\alpha,i}$, which are the filter coefficients that give the best prediction of the EGG signal from the set of microphone signals.

We can now calculate the QGG signal,

Eq. 5 $$p_t^{[1]} = \sum_{\alpha,i} x_{i,t-\alpha} \cdot c_{i,\alpha}$$

from the $c_{\alpha i}$ and microphone signals. The prediction is not at all precise because the EGG signal is a strongly nonlinear function of the glottal opening: it contains little information beyond the simple fact of whether the vocal folds are touching or not. It would be surprising indeed if one could build a linear filter that would exactly match the EGG signal.

4. EGG vs. QGG for voicing estimation.

An engineering evaluation of the QGG signal for voicing estimation is beyond the scope of the paper. Instead, we will present cases that show that (at least under some conditions) the QGG signal can provide a more reasonable estimate of the presence of voicing than either the EGG or the inverse-filtered mouth signal. This is the first, qualitative, test of the QGG.

The advantages of the QGG signal follow from its construction: because it is a linear function of the pressure near the glottis, it is well behaved during startup and shutdown of the glottal oscillator. So, unlike the EGG, it may be able to quantify partial voicing, and mark onsets of voicing precisely. Because the QGG is constructed from a time-invariant filter operating on acoustic signals, it may be more robust than algorithms based on an inverse filter (we do not discuss manual adjustment of inverse filter coefficients here, as such techniques are impractical for large speech corpora). Any time the spectral estimation step of an inverse filter fails to produce a good result, or any time the speech signal is not well represented by an all-poles transfer function, one expects the inverse-filtered signal will not reflect the glottal state. The QGG doesn't suffer from those problems.

Figure 3 and Figure 5 show examples of speech signals where the glottal oscillation is starting or stopping. The figures show that the QGG signal can sometimes provide a much better explanation of the acoustic signal than does the EGG signal. Limitations of EGG signals have previously been described elsewhere [30,31].

In Figure 3, the envelope of the QGG signal tracks acoustic power (the mouth signal), while the EGG signal shows an unnaturally sharp onset/ending. If the glottal oscillation really stopped with the EGG signal, one would have to assume a bandwidth for the first formant of only 5 Hz for the acoustic signal to persist as long as it does, which is incompatible with the known width [25,26,27] of vocal tract formants. The vocal tract simply isn't a good enough resonator for the sound to persist 30 ms after the end of glottal oscillation. Therefore, glottal oscillation must be continuing at a lower level, but without showing up on the EGG. Titze [32] and Stevens [33] have discussed this kind of small oscillation.

If we consider the physical mechanisms of the glottal oscillation and the EGG measurement process, this EGG failure is not surprising. The EGG signal measures the electrical conductivity across the glottis. During the "closed" phase of the oscillation, the vocal folds are touching in various degrees, and the conductivity provides a measure of how much they touch and how hard. However, once the glottis opens, the current between the vocal folds drops to zero, because there is an air gap between the folds. It remains essentially zero, no matter how wide or narrow the air gap. Consequently, any glottal oscillation that doesn't actually cause the vocal folds to touch won't change the electrical conductivity, and shouldn't be expected to show up on an EGG signal. We see that here; it is a common effect, showing up in low-amplitude voiced speech.

Other observations can also be explained similarly, such as the events around $t \approx 8.015\,\mathrm{s}$ in Figure 3. Imagine comparing two vowels, one uttered with amplitude just small enough so that the glottal folds don't collide, and the other uttered with slightly more amplitude so that the folds do collide on each cycle. In the two cases, the subglottal pressure is similar, and the average open area will be similar, as will the open quotient. Consequently, the total airflow per cycle past the glottis will be quite similar. If we consider a decomposition of the signal into a stack of harmonics at f_0, $2f_0$, $3f_0$, $4f_0$... the lowest harmonic will primarily measure the total air flow per cycle, and will change only gently and continuously when the glottal folds begin to collide.

However, the higher harmonics do not behave smoothly. Below the collision threshold, the oscillation is close to a simple harmonic oscillator, perhaps with weak nonlinearities resulting from the viscoelasticity of tissue and the geometry of the oscillator. Above threshold, there is a strong nonlinearity: when the vocal folds collide a simple harmonic oscillator model does not apply at all, and large amounts of power suddenly start to be generated in harmonics above the third. This is the situation that is described by the two-mass model [34,35], which typically gives a spectrum where the amplitude of the n^{th} harmonic scales as $a_n \propto n^{-1}$, or a 6db/octave slope. The harmonics now suddenly carry a substantial fraction of the total acoustic energy. Such a change can be seen in the spectra shown in Figure 4.

The effect is not confined to the low-amplitude tails of voiced sounds. For example, in Figure 5, an acoustic signal begins two periods before the first EGG activity. Again, one must assume that the vocal folds are oscillating but not yet completely closing.

Figure 6 shows a section of a low-amplitude, sustained 'o' as an extreme example where the EGG fails to explain the acoustic signal. Several times, the amplitude of the EGG signal jumps up dramatically, then drops back down a few milliseconds later. Little effect is seen in the acoustic signal, other than an increase in the power of the higher harmonics. Nor is any substantial change seen in the QGG signal. These glitches may result from a droplet of mucus intermittently forming an electrically conductive bridge between the vocal folds. Alternatively, we could be seeing an oscillation where the vocal folds come within a whisker of touching, and some small perturbation briefly increased the amplitude or reduced the spacing just enough to make them collide. The noteworthy observation here is that the QGG signal is a much better predictor of the acoustic signal than the EGG.

These problems we have displayed are not hard to find, occurring at these levels in 6 of 304 voiced syllables inspected. Because the problems seem to be associated most with glottal oscillations where the vocal folds do not contact, we expect that EGG problems should be much more common in languages that make more extensive use of murmurs (*i.e.*, a 'breathy' voice quality), most notably Hindi. The inverse filtered signal also tends to behave badly for low amplitude voicing or other conditions where the power in the higher harmonics is very low. Among the displayed signals, the QGG provides signals which display a strong contrast between voiced and unvoiced regions, and have most of their power in the fundamental to reduce the likelihood of octave errors in any following pitch tracker.

5. QGG as a measure of amplitude/emphasis

The Quasi-glottogram signal is valuable for more than correcting voicing errors. It also provides an estimate of the amplitude of the oscillatory flow through the glottis. We expect that this amplitude will be a better predictor of prosodic emphasis and a better measure of the speech effort being expended by the speaker than is the total acoustic power of the mouth signal or the inverse filtered mouth signal or EGG.

Amplitude has been known to be a significant component of prosody since the 1950s [36,37,38,39,40], and into more recent literature [41,42,43,44,45]. However, all these studies have been severely limited by the large intrinsic variability of speech amplitude measurements. The experimental designs (*e.g.*, ANOVA analysis on p. 190 of [41]) invariably compare the prosodic effect in question to the unpredictable variations. Reducing this variability can be seen to be just as good as having a larger effect to measure. This is one value of the QGG: it allows a cleaner, low variance amplitude measurement, and should lead to more conclusive experiments.

5.1. Model of amplitude variance

To justify our intuition that the QGG signal will allow better amplitude measurements, consider a toy model of the speech apparatus: a glottal source that drives the vocal tract, which we model as a time-varying filter. Loosely speaking, the variability of the amplitude outside the mouth comes from two sources: intrinsic variability in U, and changes in the coupling through the vocal tract transfer function, V. Since the two variances add, the variability of the mouth power will be greater than the variability of glottal power. Consequently, we expect that the best linear estimators of the glottal source should have the lowest variability. We can use this as a figure of merit to compare algorithms: less variable estimators are better and probably closer to the glottal signal.

We will work in a short-time Fourier transform basis to conveniently describe speech-like signals. Signals are then indexed with two parameters: a time index, t, which locates the transform's window, and a frequency index, ω, for the low-resolution spectrum in the window. In this toy model, the glottal source changes its amplitude but not its spectral shape: $U(t,\omega) = f(t)\cdot g(\omega)$, where we assume that the amplitude, $f(t)$, changes slowly compared to structure in the glottal spectrum, $g(\omega)$. We can safely assume $\sum_{\omega}\left|g(\omega)\right|^2 = 1$ without affecting the model, as the overall amplitude can go into $f(t)$.

Next, we can write a time- and frequency- dependent transfer function for the vocal tract: $h(\omega,\phi(t))$, where $\phi(t)$ is the vocal tract configuration (roughly, the phone) at time *t*. The pressure outside the mouth is then $s(\omega,t) \approx f(t)\cdot g(\omega)\cdot h(\omega,\phi(t))$, and we can sum over frequency to get the RMS amplitude of the mouth signal: $a^2(t) = \sum_{\omega}\left|s(\omega,t)\right|^2 = f^2(t)\cdot z^2(\phi(t))$, where $z^2(\phi) = \sum_{\omega}\left|g(\omega)\cdot h(\omega,\phi)\right|^2$ shows how efficiently power is coupled from the glottal source out through the mouth for a particular phone ϕ.

We can now take the log of the power to write

Eq. 6 $\log(a(t)) = \log(f(t)) + \log(z(t))$.

As long as variations of the vocal tract are uncorrelated with changes in the larynx, the variances of the two right hand terms add, and we can conclude that $\mathrm{var}(\log(a)) = \mathrm{var}(\log(f)) + \mathrm{var}(\log(z))$. Since $\mathrm{var}(\log(z)) > 0$, $\mathrm{var}(\log(a)) > \mathrm{var}(\log(f))$. In other words, the amplitude outside the mouth is always (in this toy model) more variable than the amplitude of *U*. The same conclusion follows if you consider *h* to include the vocal tract plus an arbitrary linear operator: thus any filtered version of the mouth signal will still be more variable than *U*.

The limits to the assumption that *f* and *z* are uncorrelated come from two sources: First, the speaker's intentional muscle motions can have correlations between the glottis and the vocal tract (*e.g.,* a hypothetical language might specify that high vowels are always spoken in a pressed voice). Second, some vocal tract configurations with tight constrictions can change the glottal waveform. However, neither circumstance seems common.

As a concrete example of this, consider a vowel where f_0 matches the first formant frequency, f_1. Acoustic power will then be efficiently coupled from the glottis out the mouth, because a peak of $g(\omega)$ matches with a peak of $h(\omega,\phi(t))$, z will be large, and the amplitude at the mouth will therefore be large. On the other hand, if $\frac{3}{2}f_0 = f_1$ with the same amplitude of U, the fundamental frequency will be below the first resonance of the vocal tract, z will be small, and the amplitude at the mouth will be small. Comparing the two cases, we see that substantial variance in amplitude can be generated as the sound wave propagates through different configurations of the vocal tract.

This toy model contains several loose assumptions and shouldn't be taken too far, but it does give important clues for finding good algorithms, since the mathematics remains valid if the transfer function, *h* includes the behavior of the microphones and a linear signal processing algorithm. For example if *h* is time-invariant, var(log(*z*)) will be zero, and the variability of the amplitude estimate, var(log(*a*)), will be as small as possible. Conversely, if *s* is always close to *g*, the transfer function must be near unity, so that var(log(*z*)) must be small and thus var(log(*a*)) will be as small as possible. So, we expect that algorithms that are good glottal estimators will give stable amplitude estimates and vice versa. Note that for further arguments, we do not require any of the details of the toy model, merely this conclusion, which is independently testable, and likely to be true independent of the model.

Purely pragmatic considerations will also lead us to the same figure of merit. If one is studying prosody, then any variation of amplitude that is a function of the phoneme should be considered noise: it prevents one from comparing prosodies of different words. Good comparisons are only possible if the amplitude measure is independent of phoneme. So, we would like to improve the signal-to-noise ratio of prosody experiments by reducing the noise, which again means finding an amplitude estimator that is less variable.

5.2. Comparison of QGG & acoustic power variance

We conducted a second, quantitative test. This test directly establishes the usefulness of the QGG signal as a measure of amplitude prosody. We also show that the QGG signal allows a very steady estimation of amplitude, less variable than the result of other standard linear estimators. Following the logic in Section 5.1, this test indirectly establishes that the QGG signal is a reasonable estimator of a filtered version of U.

We used the database of 979 utterances in the form *"Ta shuo X san tian."* , described above. We calculated the QGG for all the utterances, along with an inverse-filtered **M**, low-pass filtered **M**, and the raw **M**. The boundaries of the variable syllable (X) were hand-segmented, and an algorithm (ESPS/waves *get_f0*) was run to find the two voiced regions on either side of the segmented area, and also the voiced region inside X. Four of the utterances were voiced through between "*ta*" and "*shuo*" and were dropped. We then calculated the mean power near the center of the five voiced regions, using a cosine window.

Table I shows the standard deviation of log(power) for each combination of region and signal. In every case, the QGG signal is more reproducible than the others, yielding (on average) a standard deviation 46% smaller than the corresponding low-pass filtered speech, 38% smaller than unprocessed speech, and 20% less than the inverse filtered signal. These improvements in SNR are conservative limits, as the speech contains some intrinsic variability that cannot be removed by signal processing.

Comparisons of cells in Table I have several implications. First, one expects the variation of the frame (regions 1, 2, 3 and 4) to be small, where we always have the same syllable in the same position. In contrast, the variation in the X region should be bigger, because the syllable identity changes. This is reflected in Table I, where the X region shows the largest variation under all conditions.

Why is the QGG variability not smaller in the X region, if it indeed removes the effect of changes in the vocal tract? It is not smaller because, unfortunately, the QGG is not an estimator for *U*, but rather (see section 3) a filtered version of *U*, $q \cdot U$, where the filter, *q*, is time-invariant but can vary as a function of frequency. The QGG estimate can therefore vary with the f_0 of the speech.

We tested this by choosing a subset of the data that all has similar f_0, and re-calculating the standard deviation of region X. We chose all 260 syllables that have Mandarin tone 1. This is a high, level tone, which is the same tone that occurs in regions 1, 2, 3, and 4. In this subset of the data, the f_0 trajectory is relatively flat in each syllable, as well as across the whole utterance. The results are displayed in column X' of Table I. The standard deviation of the amplitude of all the signals is lower, but the QGG drops most dramatically.

One can also see both the pitch-dependence of the QGG and its relative insensitivity to the vocal tract configuration in Figure 7. This is a scatter plot of the mean f_0 of all syllables *vs.* the measured amplitude of the QGG signal. We used the same window (as defined above) to calculate the mean f_0 as was used for the amplitude measurement. Syllables with tone 1 are seen as a tight cluster in the upper left corner. That cluster spans the full range of phonemic variation, covering all vowels in combination with a variety of consonants and glides. The other syllables in Mandarin (shown as dots) have either low pitch, or they are rising from or falling to low pitch, thus they have an average pitch below that of tone 1. There is a clear trend of increasing amplitude measurement with increasing pitch, presumably at least partially as the result of *q*. One can empirically correct for this trend, bypassing the QGG signal through a properly defined filter, if one knows how much of the effect is the result of *q*, and how much is the result of the speaker's glottal flow changing as a function of pitch.

While the QGG algorithm dramatically reduced the variation of Regions 1 and 2 (to around 10% in table I), it was less successful in Regions 3 and 4. This suggests that there is more inherent variability in Regions 3 and 4, which may well be some carry-over from the pitch and phoneme change in region X. This larger inherent variation can also be seen for all the signals, although not as clearly because the other algorithms don't yield as reproducible an amplitude signal as the QGG.

5.3. QGG for amplitude prosody

As a further test with the same database (Table II), we attempted to eliminate any changes in amplitude by predicting the amplitude of each region in terms of the amplitudes of the other regions. Essentially, this normalizes the measured amplitudes to the rest of the utterance, and would eliminate the effect of a uniform change in amplitude from utterance to utterance. The goal here is to reduce the variability deriving from the experimental subject (*e.g.,* from changes in the volume of inspired air) and focus more tightly on variations that result from the signal processing.

We fit a least-squares linear predictor to the logs of the amplitudes, *e.g.,*

Eq. 7 $$\log(\hat{A}_X) = b_0 + b_1 \cdot \log(A_1) + b_2 \cdot \log(A_2) + \mathrm{K} \ ,$$

and then measure the RMS size of the residual, $\log(A_X) - \log(\hat{A}_X)$. Here, A_R is the amplitude of the R^{th} region (a RMS average of the signal inside a cosine window), and $\hat{A}$ is the predicted amplitude, based on the other regions. The predictor is a five-parameter linear fit, and we fit separate predictors for regions 1, 2, X', 3, and 4. Again, the QGG leads to a nicely small variance: it has predictable amplitude. Surprisingly enough, the low-pass filtered signal is comparably predictable, even though it's performance before prediction (Table I) is quite poor.

The most important terms are generally those that predict A_3 in terms of A_4 and vice-versa. After prediction, the frame regions on both sides of X have similar variabilities, approximately 8% for the QGG signal. This remainder seems intrinsic to the speaker. The excess variation in regions 3 and 4 is gone, even though the variation in regions 1 and 2 is practically unchanged. This suggests that amplitude variations in X, which are driven by phoneme and f_0 changes, carry forward into the following syllables.

Because region X' contains a diverse set of different syllables while the frame regions (regions 1, 2, 3, 4) always have the same syllable in the same position, we expect more variability in the amplitude of X'. Tables II shows this. However, changing syllables only disturbs QGG amplitude measurements by 11% (beyond the intrinsic 8%), so long as the pitch is reasonably stable.

6. Summary

We have shown that the QGG algorithm can produce a useful, noninvasive estimate of the glottal flow. When used to analyze speech, it can be well behaved under conditions where EGG and inverse-filtered signals make gross voicing errors. It also yields substantially more stable amplitude measurements than other techniques. The QGG algorithm should find applications in studies of the amplitude part of prosody. We also see applications in text-to-speech systems, where there is a need for reliable automatic processing of speech data, and possibly in medical screening or diagnostics of voice disorders.

We acknowledge Robert Kubli and Gary Elko for equipment and advice, Mohan Sondhi and Joseph Hall for discussions and Ronald J. Baken and three reviewers for substantial comments.

Tables

Fractional standard deviation of amplitude	**Region 1** ***ta***	**Region 2** ***shuo***	**X** ***X***	**X'** **(tone 1 only)**	**Region 3** ***san***	**Region 4** ***tian***
Raw	0.22	0.16	0.48	0.32	0.31	0.29
Low-pass Filtered	0.29	0.23	0.42	0.34	0.32	0.35
Inverse Filtered	0.14	0.14	0.42	0.24	0.22	0.24
QGG	0.09	0.09	0.37	0.16	0.19	0.20

Table I: Variability of speech amplitude, after processing by four algorithms. Region X is the variable syllable, and shows larger variability because the amplitude is a function of the phoneme in X.

Fractional standard deviation of amplitude, after linear prediction.	**Regions 1 & 2** ***ta shuo***	**X'** ***(tone 1 only)***	**Regions 3 & 4** ***san tian***
Raw	0.16	0.32	0.11
Low-pass Filtered	0.09	0.16	0.06
Inverse Filtered	0.14	0.23	0.10
QGG	0.08	0.14	0.09

Table II: Unpredictable variability of speech amplitude, after processing by four algorithms. Here, we use the subset of the data where X has tone 1 (a high level tone), so that the pitch matches regions 1, 2, 3, 4. The amplitude in each region was predicted in terms of the other four regions; the table shows the remainders.

Figures

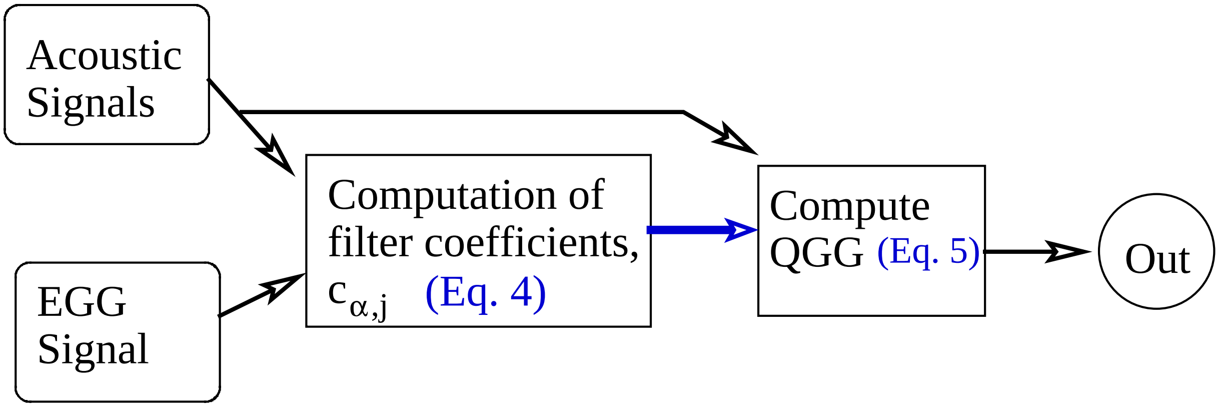

Figure 1. Diagram of the data and signal flow in the computation of the QGG.

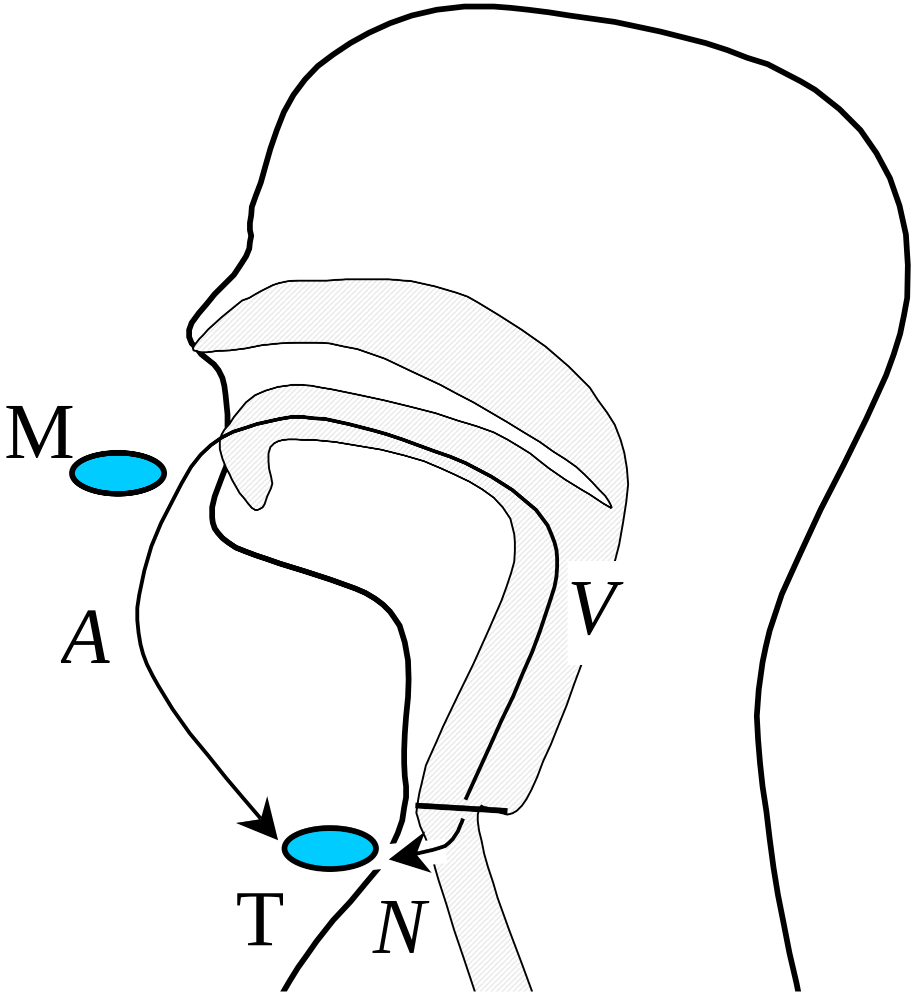

Figure 2. Schematic of microphone placement and signal paths.

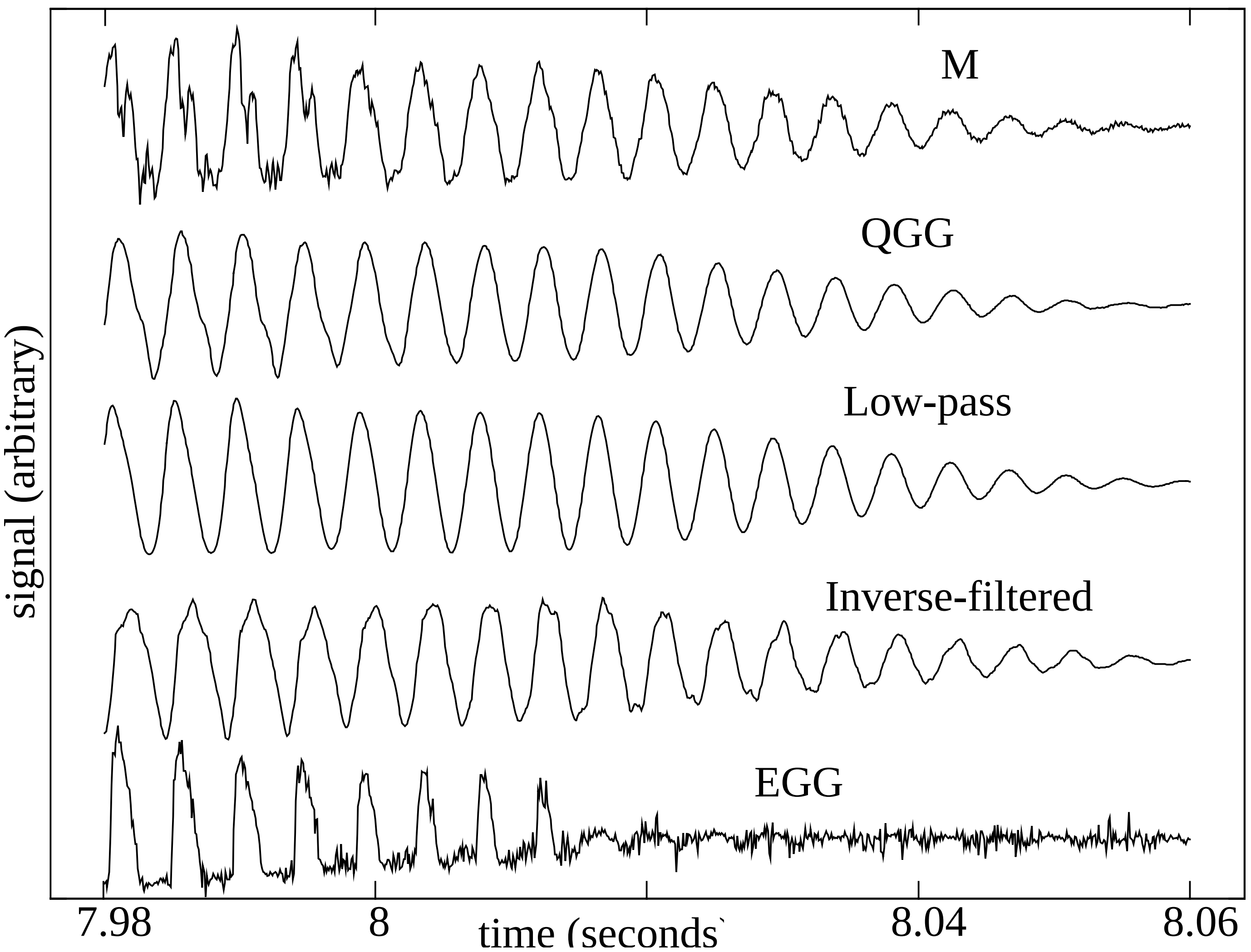

Figure 3. Comparisons of estimators of the glottal waveform. The signal is the off-glide [aɪ] in “high”. The EGG stops more than 30ms before the end of the glottal oscillation.

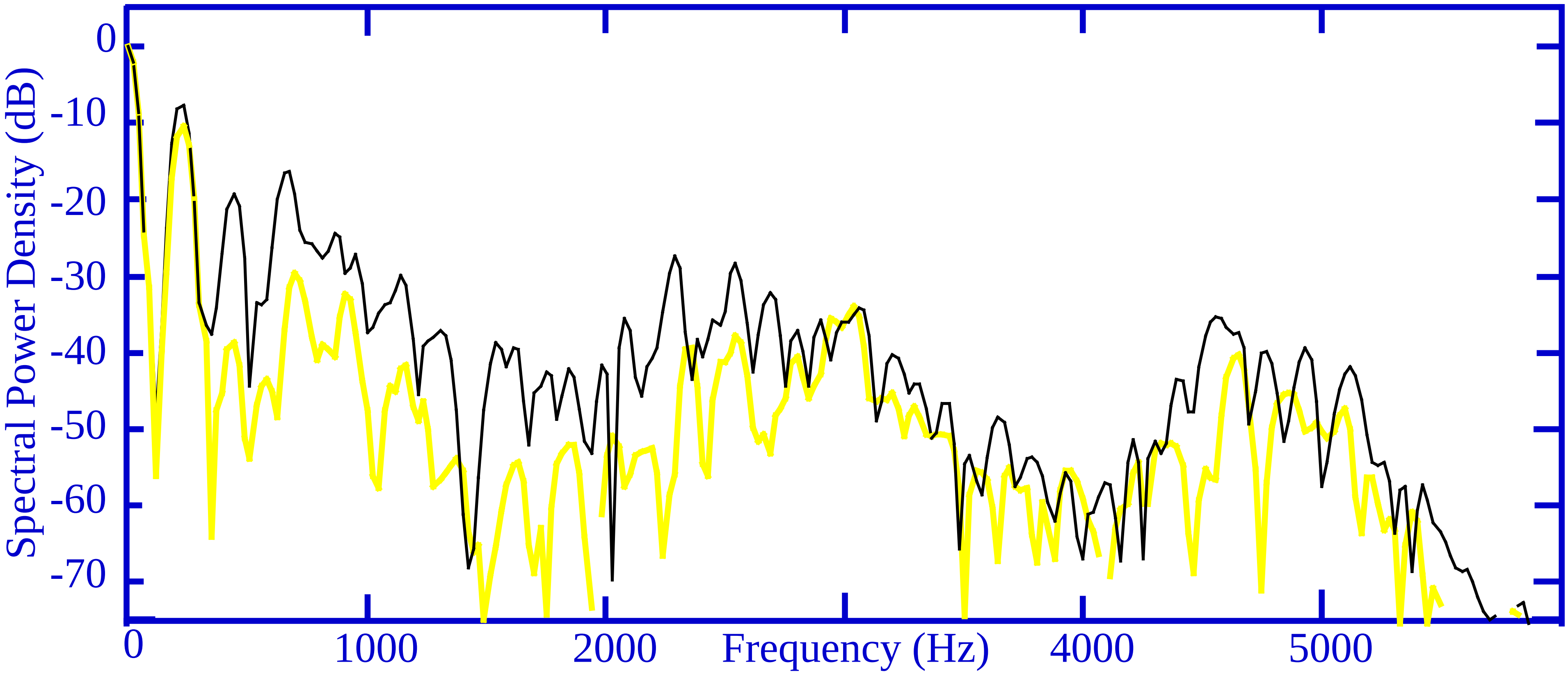

Figure 4. LPC spectra (16 ms windows, from AR(14) model) on either side of 8.015 s in Figure 3. The think, black curve is before 8.015s, when the glottal folds are colliding, and the wide, grey curve is after. The fundamental (225Hz) is essentially unchanged in amplitude, but the power in the harmonics drops by about 10 dB.

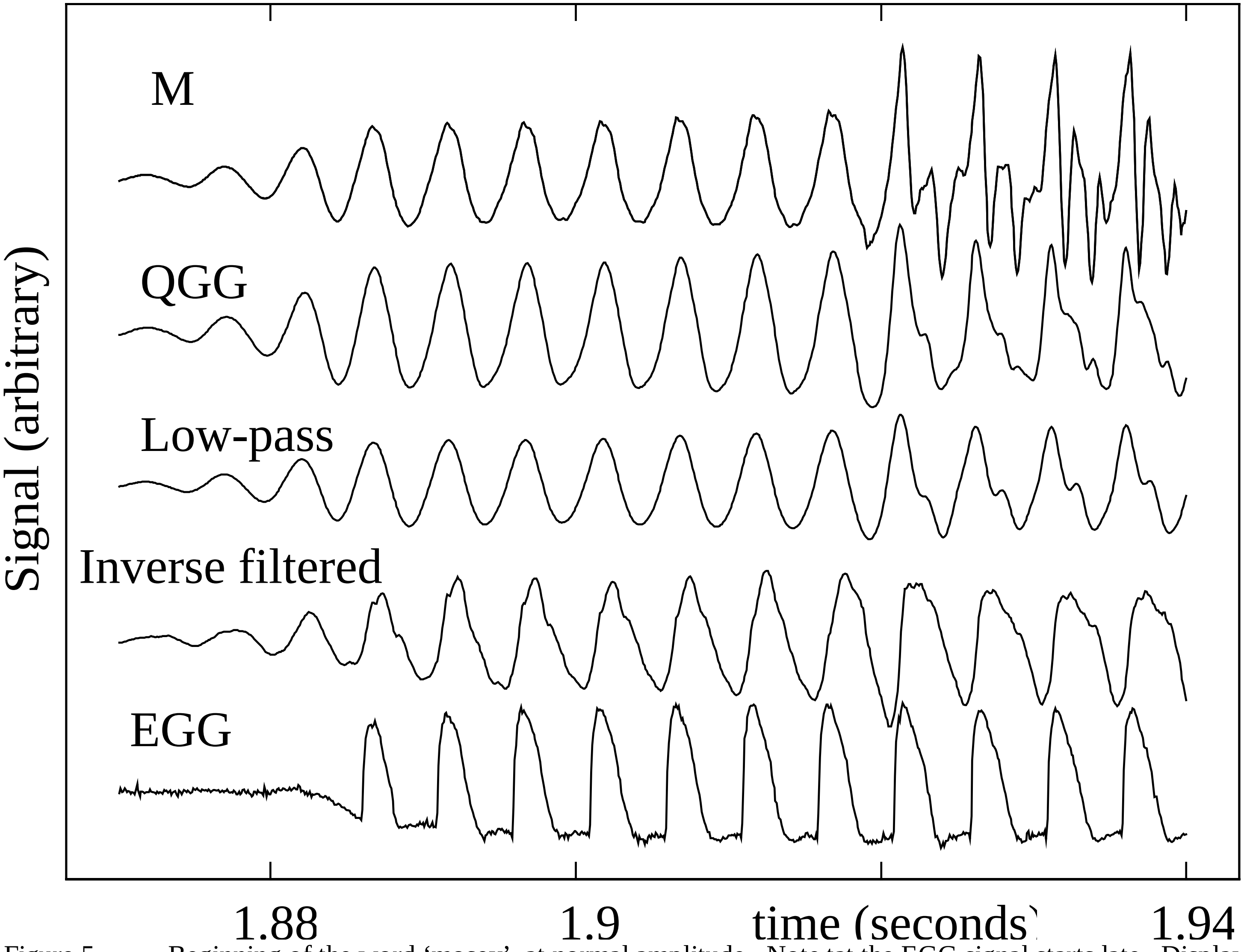

Figure 5. Beginning of the word ‘mosey’, at normal amplitude. Note tat the EGG signal starts late. Displayed as in Figure 3.

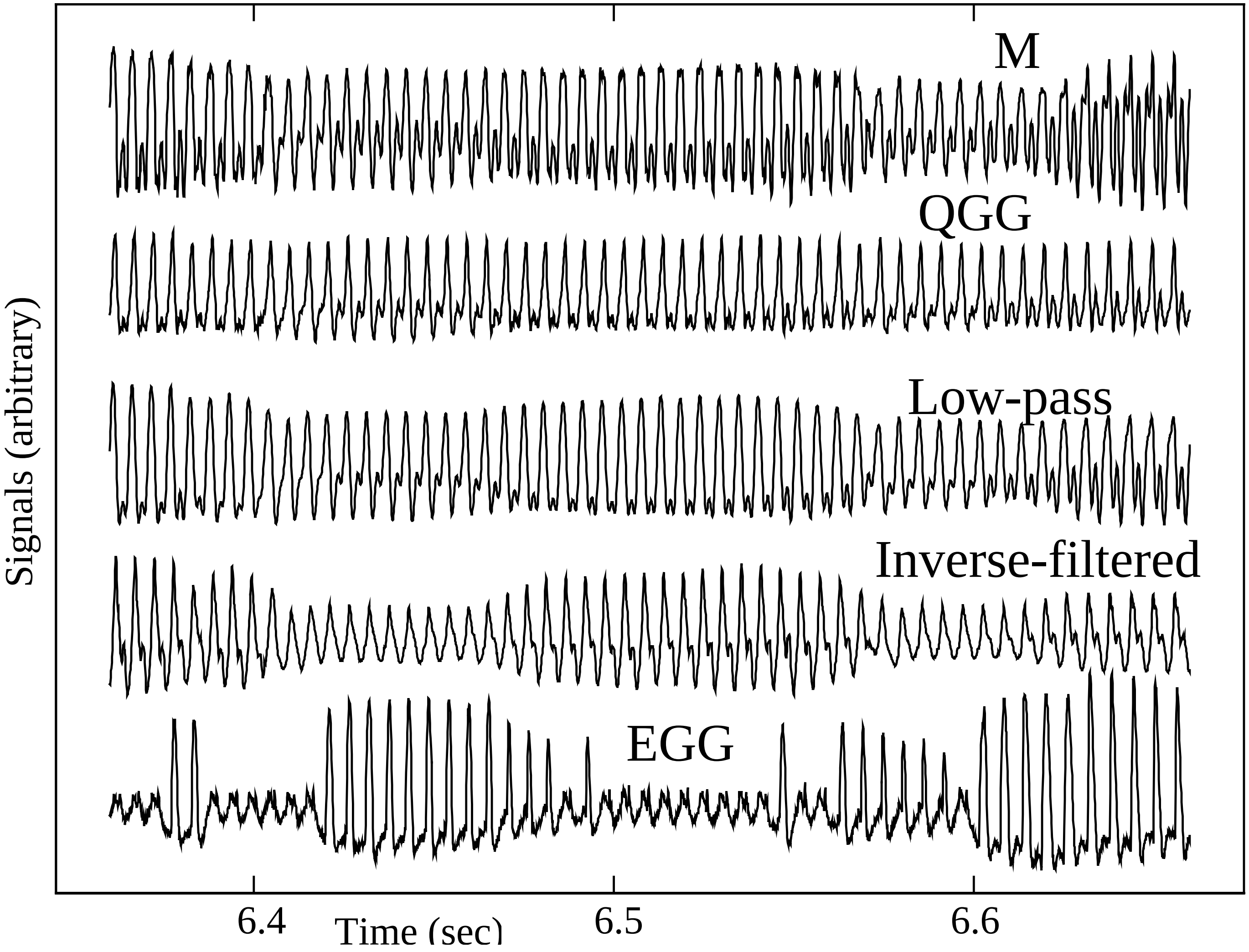

Figure 6. The middle of sustained low amplitude ‘o’ phonation showing major EGG changes (bottom) without large changes in the speech signal (top). The various signals are labeled.

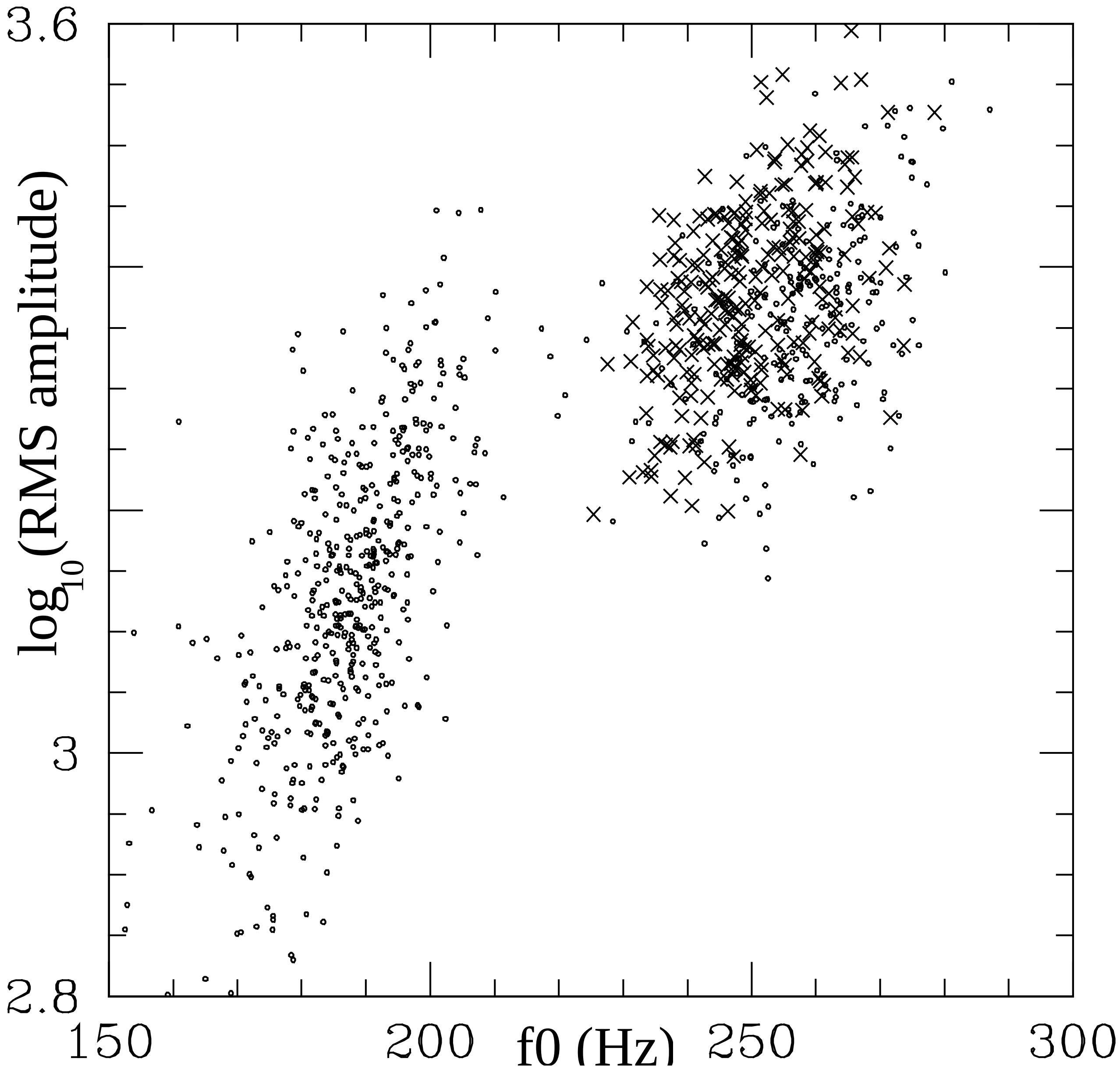

Figure 7. Amplitude measurement vs. mean pitch of a syllable for QGG signal. Syllables with tone 1 (used in column X' of Table I and Table II) are shown as 'x'. Other tones are shown as dots. In this plot, the pitch-dependencies (from one tone to another) are larger than the dependence on the syllable (within the tone 1 syllables).

References

[1] Sondhi, M. M., 1975, “Measurement of the Glottal Waveform,” *J. Acoustical Soc. Am.*, 57, 228-232.

[2] Baken, R. J. and Orlikoff, R. F., 2000, *Clinical Measurement of Speech and Voice,* Singular Publishing Group, ISBN 1-56593-869-0, p. 431-432 and references therein.

[3] Baken, R. J. and Orlikoff, R. F., *op. cit*., p.311-322 and references therein.

[4] Farnsworth, D. W., 1940. “High-speed Motion Pictures of the Human Vocal Cords,” *Bell Laboratories Record,* 18, pp. 203-208.

[5] Baken, R. J. and Orlikoff, R. F., *op. cit*., pp. 394-406 and references therein.

[6] Sonneson, B. (1960). “On the anatomy and vibratory pattern of human vocal cords with special reference to a photo-electrical method for studying the vibratory movements,” *Acta Oto-Laryngologica,* Supplementum **156**, 1-81.

[7] Gerratt, B. R., Hanson, D. G., Berke, G. S., and Precoda, K. (1989). Phottoglottography: A clinical synopsis, J. Voice, **5**(2), 98-105.

[8] Baken and Orlikoff, *op. cit*., pp. 322-324 and references therein.

[9] Baken and Orlikoff, *op. cit*., pp. 413-427 and references therein.

[10] Fabre, P. 1957. “Un procédé électrique percutané d’inscription de l’accolement glottique au cours de la phonation: glottography de haute fréquence. Premiers résultats.”, *Bull. Acad. Natl. Med.* **141**, 66-69.

[11] Orlikoff, R. F., 1991, “Assessment of the Dynamics of Vocal Fold Contact from the Electroglottogram: Data from Normal Make Subjects”, *J. Speech and Hearing Research*, **34**, 1066-1072.

[12] Sonesson, B. (1960) “On the anatomy and vibratory pattern of the human vocal folds,” *Acta Otolaryngologica*, **Suppl. 156**, 1-80.

[13] Timcke, R., von Leden, H., and Moore, P. (1958). “Laryngeal vibrations: Measurements of the glottic wave: Part 1. The normal vibratory cycle,” *Archives of Otolaryngology*, **68**, 1-19.

[14] Baer, T., Löfqvist, A., and McGarr, N. S. (1983). “Laryngeal vibrations: A comparison between high speed filming and glottographic techniques,” *J. Acoustical Soc. America*, **73**, 1304-1308.

[15] Baer, T., Titze, I. R., and Yoshioka, H. (1983). “Multiple simultaneous measures of vocal fold activity,” in D. M. Bless and J. M. Abbs (Eds.), *Vocal Fold Physiology: Contemporary research and clinical issues* (pp. 229-237). San Diego, CA: College Hill Press.

[16] Gilbert, H. R., Potter, C. R., and Hoodin, R. (1984). “Laryngograph and a measure of vocal fold contact area,” *J. Speech and Hearing Research*, **27**, 556-565.

[17] Baken and Orlikoff, *op. cit*., pp. 433-436 and references therein.

[18] Ishizaka, K., Matsudaira, M., and Kaneko, T, 1976. “Input acoustic-impedance measurement of the subglottal system,” J. Acoust. Soc. Am. 60:1, pp. 190-196.

[19] Milutinović, Z., Mijić, M., Djurica, S. (1997). “Activity of the sobglottic voice (‘chest resonator’): an echo-tomographic and acoustic study,” Eur. Arch. Otorhinolaryngol 254: 292-297.

[20] K. Stevens (1998), *Acoustic Phonetics*, MIT Press, ISBN 0-262-19404-X. pp. 66-68, 76, 94.

[21] Laryngograph, Ltd. 1 Foundry Mews, Tolmers Square, London NW1 2PE, U.K.

[22] Entropic Research Laboratory. Code attributed to Brian Sublett, John Shore, David Talkin, and Derek Lin. It does a 14th order LPC analysis with a Hanning window, with a pre-emphasis coefficient of 0.97.

[23] K. Stevens (1998), *op. cit .*, p.200-202.

[24] Weider, Sol, 1973. *The Foundations of Quantum Theory,* Academic Press, New York, pp. 44-59.

[25] Fant, G. (1962). “Formant Bandwidth Data”. *Speech Transmission Laboratory Quarterly Progress and Status Report **2-3***, Royal Institute of Technology, Stockholm, 1-3.

[26] Fujimura, O. and Lindqvist, J. (1971). *Sweep-tone measurements of vocal tract characteristics.* J. Acoustical Society of America, **49**, 541-558.

[27] K. Stevens *op. cit.* p. 152-167, 264

[28] K. Stevens, *op. cit.*, p. 196-198 and references therein.

[29] Press, W. H., Teukolsky, S., Vetterling, W. T., Flannery, B. P., *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edition, Cambridge University Press, 1992, ISBN 0-521-43108-5. p. 564-565.

[30] Colton, R. H., Conture, E. G., (1990). “Problems and pitfalls of electroglottography,” *J. Voice* **4,** pp. 10-24.

[31] Hanson, D. G., Gerratt, B. R., Berke, G. S., (1990). “Frequency, intensity, and target matching effects on photoglottic measures of open quotient and speed quotient,” *J. Speech Hearing Res.* **33**, pp. 45-50.

[32] Titze, I. R. (1988). “The physics of small amplitude oscillation of the vocal folds,” *J. Acoust. Soc. Am.* **83(4)**, 1536-1552.

[33] K. Stevens, *op. cit.* p. 91.

[34] Ishizaka, K. and Matsudaira, M. (1968). "What makes the vocal cords vibrate?", In *Proceedings of Sixth International Congress of Acoustics*, Tokyo, B1-3.

[35] Ishizaka, K. and Flanagan, J. L. (1972). "Synthesis of voiced sounds from a two-mass model." *Bell System Technical Journal*, **51**, 1233-1268.

[36] Fry, D. B., (1955), "Duration and intensity as physical correlates of linguistic stress," *J. Acoustical Soc. Am.* **30**, 765-769

[37] Fry, D. B. (1958), "Experiments in the perception of stress," *Language and Speech*, **1**, 126-152.

[38] Bolinger, D. L. (1958), "A theory of pitch accent in English," *Word*, **14**, 109-149.

[39] Lieberman, P. (1960), "Some acoustic correlates of word stress in American-English," *J. Acoustic Soc. Am.*, **32**, 451-454.

[40] Hadding-Koch, K. (1961), *Acoustico-Phonetic Studies in the Intonation of Southern Swedish*, C. W. K. Gleerup, Lund, Sweden.

[41] Pollock, K. E., Brammer, D. M. and Hageman, C. F. (1990), "An acoustic analysis of young children's productions of word stress", *J. Phonetics*, **21**, 183-203.

[42] Kehoe, M., Stoel-Gammon, C., and Buder, E. H. (1995), "Acoustic Correlates of Stress in Young Children's Speech", *J. Speech and Hearing Research*, **38**, 338-350.

[43] Sereno, J. A., Jongman, A. (1995), "Acoustic Correlates of Grammatical Class," *Language and Speech* **38(1)**, 57-76.

[44] Beckman, M. (1986), *Stress and non-stress accent*, Dordrecht: Foris Publications.

[45] Sundberg, J., Elliot, N., Gramming, P., and Nord, L. (1993) "Short-term variation of subglottal pressure for expressive purposes in singing and stage speech – a preliminary investigation," *J. Voice* **7(3)** 227-234.