

INVESTIGATING SPEAKER FEATURES FROM VERY SHORT SPEECH RECORDS

Brian LaRoy Berg and A. A. (Louis) Beex

Systems Group - DSP Research Laboratory
The Bradley Department of Electrical Engineering
VIRGINIA TECH
Blacksburg, VA 24061-0111

ABSTRACT

A procedure is presented that is capable of accurately extracting various speaker features, and is of particular value for analyzing records containing single words and shorter durations of speech. By taking advantage of the fast convergence of adaptive filtering, the approach is capable of modeling the nonstationarities due to both the vocal tract and vocal cord dynamics. This procedure is quite simple, requires no manual intervention, and is particularly unique because it derives both the vocal tract and glottal signal estimates directly from the time-varying filter coefficients rather than the prediction error signal. Several glottal signals are derived using this procedure, and are plotted to demonstrate the kind of glottal characteristics obtained therein. Finally, in order to provide a more quantitative performance measure, the procedure is used in a simple automatic speaker identity verification application.

1. INTRODUCTION

Features found within short speech records are important for modeling the characteristics of a speaker's voice. Such speaker features are valuable in applications such as speech coding and synthesis. In general, these speaker features are both behavioral and inherent. Inherent (or anatomical) features depend upon the anatomy of the vocal cord and vocal tract (i.e., air passages above the vocal cords). The vocal tract anatomy refers to the size and shape of the vocal tract and is determined by the utterance (which is a function of mouth and tongue position) and to some degree the sex, size, and age of the speaker. Formants (i.e. speech resonances) are common features used to define the vocal tract. The vocal cord anatomy determines the pitch, breathiness, and vocal register (e.g. creakiness) of the speaker. Examples of behavioral features are dialect and voice expressiveness and these relate to the longer-term dynamics of the speaker's vocal tract (particularly, the movement of the tongue and jaw) and vocal cords.

Various approaches have been used to extract speaker features. The most popular speech analysis algorithms use linear prediction (such as the Autocorrelation or Covariance techniques) to extract formants as they vary over time. The vocal cord features remain within the residual signal, which is essentially the prediction error of the analysis algorithm. Speech coding schemes have parameterized this information in various ways; two popular approaches

The first author was supported throughout parts of this work by a Bradley Fellowship and summer internship programs at Bellcore and Bell Atlantic. He is now with Hewlett-Packard Company, 5601 Lindero Canyon Rd, Westlake Village, CA 91362.

are known as multipulse excited linear prediction and code excited linear prediction.

In pursuit of better models, speech synthesis researchers have developed an analysis approach known as glottal inverse filtering (GIF) [2, 4]. GIF uses short linear prediction analysis frames, which are centered over a specific region within each pitch period to minimize inaccuracies due to speech nonstationarities. From this accurate linear prediction model, a residual signal is computed to represent what is known as the differential glottal waveform. This glottal waveform estimate has been compared to results obtained from a physical analysis of the vocal cords [2] and with older inverse filtering techniques [5] and has been found consistent therewith. It has also been adopted as the source signal for popular speech synthesis systems known as formant synthesizers [5, 6].

This paper investigates the speaker features that exist in short speech records, of word length and less, and presents an analysis algorithm that requires no manual intervention, to accurately extract and efficiently model them. The proposed procedure is based on standard adaptive (i.e. recursive) filtering and thus computes a formant estimate at every sample, rather than once per frame [3]. Hence the operations required for this analysis procedure consist of an adaptive filter, inverse filter, pitch detector, voiced/unvoiced (v/uv) detector, endpoint detector, and pre-emphasis; all of which are quite common. Furthermore, simple detection routines (v/uv, endpoint and pitch) are adequate for successful operation.

Records from several speakers are analyzed to demonstrate the benefits of this method for modeling fine speech detail; even the vocal cord dynamics that produce each pitch period. An example of how this procedure can be used successfully for speaker identity verification is also shown.

2. ADAPTIVE FORCED RESPONSE INVERSE FILTER APPROACH

To demonstrate the recursive filtering operation on speech, Figure 1 shows time-varying spectral estimates over a single speech period. These 66 speech samples were obtained by digitally recording a male senior (who will be referred to by the letters WES) using the system described in the appendix. These samples correspond to samples 45 to 110 of the waveform in Figure 2. As with standard LPC analysis, the adaptive filter uses a 10th order autoregressive (AR) model.

The high temporal resolution provided by these time-varying spectral estimates allows continuous tracking of the formants as they are affected by the vocal cord behavior. Note that the first 25

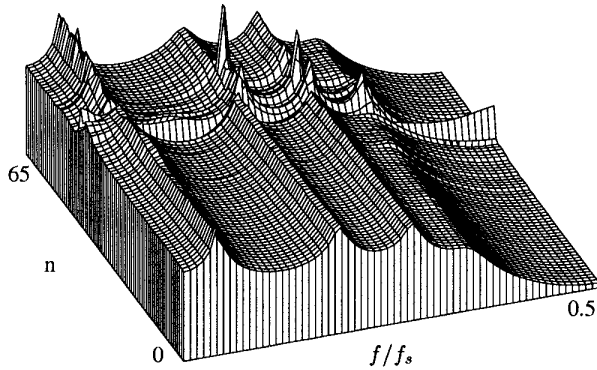


Figure 1: Instantaneous Spectra From RLS.

to 30 estimates are quite stationary. In the next 10 to 15 samples the formants dampen and broaden until the 35th estimate where the formants are excited, and the bandwidths suddenly decrease. For the following 5 to 10 samples, the energy in the high frequency formants increases, but decreases in the lowest frequency formant.

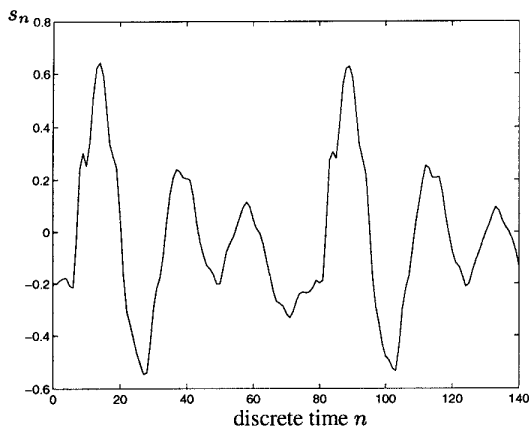


Figure 2: Speech Record Used For Illustration.

These events correspond with other studies which have gone a step further by relating them to the specific mechanics of the vocal cords. The 35th sample at which the formants are excited is the key event, because it results in the generation of the next speech period as shown in Figure 2 starting at $n = 80$. This is known as the main excitation and corresponds to the moment of glottal closure [4]. A well-known consequence of this event is the strong excitation of the high harmonics that was observed for the following 5 to 10 samples [6]. So accordingly, the 10 to 15 sample interval prior to the 35th sample should correspond to when the glottis is open. The observed time-varying frequency response certainly confirms this assumption as other studies have also found that glottal opening indeed tends to cause a damping of the lower frequency harmonic [9].

In summary, the above observations on vocal cord behavior complement those observed in other studies. These events are well understood due to the popularity of the glottal waveform (resulting from the introduction of glottal inverse filtering as an alternative

to expensive physical glottal waveform measurements). However, the significance of the analysis made here is that, unlike in the standard glottal inverse filtering technique, the above observations of glottal events were made using the prediction coefficients rather than the prediction error signal. The following sections describe how to separate the vocal tract and vocal cord information from these time-varying spectral estimates.

2.1. Adaptive Vocal Tract Modeling

As observed in Figure 1, the recursive filter analysis algorithm provides several linear predictive formant estimates over each pitch period. In order to accurately model the vocal tract without the excitation effects, a linear prediction parameter vector should be chosen from within the closed glottis interval (CGI), as is done in glottal inverse filtering. Since vocal tract characteristics vary relatively slowly, it is not necessary to detect the closed glottis intervals and store the vocal tract features for every pitch period. Hence the proposed approach avoids the difficult task of accurate pitch detection by using a time-domain feature extraction routine to identify only the obvious main excitations, of which only one vector, say \mathbf{a}_{n_c} , is chosen within each successive 30 ms analysis frame.

2.2. Forced Response Inverse Filter for Vocal Cord Modeling

Figure 1 demonstrated how the prediction coefficients obtained from the recursive analysis algorithm contain a description of the glottal behavior. A procedure is described here that aims to model this behavior using a 1-dimensional signal rather than by analyzing the variations of the p prediction coefficients.

The first step of the procedure is to obtain the step response of the time-varying filter defined by the prediction coefficients:

$$h_n = u_n + \sum_{i=1}^p a_{n,i} h_{n-i}.$$

Finally, to model how the statistics of this signal change over a pitch period as a result of the vocal cord behavior, the step response is then injected into the time-invariant inverse filter defined by the prediction coefficient vector \mathbf{a}_{n_c} , obtained in Section 2.1:

$$g_n = h_n - \sum_{i=1}^p a_{n_c,i} h_{n-i}.$$

The overall flow diagram of the proposed procedure is given in Figure 3.

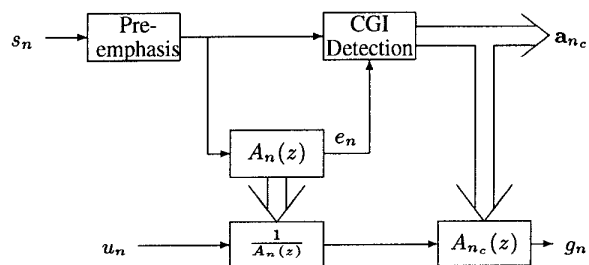


Figure 3: Flow Diagram of The Proposed Analysis Procedure.

Note that since the previous time-invariant filter represents the inverse of the vocal tract model for the corresponding period (which is assumed to be stationary), the output should be free from vocal tract information. If the speech signal itself is stationary over the particular pitch period (which it never is), the output of the inverse filter will be unity. Results different from unity within the pitch period thus reflect the nonstationarities associated with the vocal cord operation. Hence the inverse filter output $\{g_n\}$ will be referred to as the glottal signal. For example, Figure 4 shows the glottal signal estimate for the speech period reflected in Figure 2.

While the intent of this aspect of the proposal is to present a straightforward approach for extracting glottal features rather than to propose a new approach for computing the glottal waveform, it is interesting to observe its similarities and differences with GIF. Perhaps the most obvious point is how they both use an inverse filter of prediction coefficients extracted from the CGI. However, recall that GIF techniques use the speech waveform as the signal to be filtered, whereas the proposed approach uses the step response which is basically the speech signal with much of the random component replaced by the unit step signal.

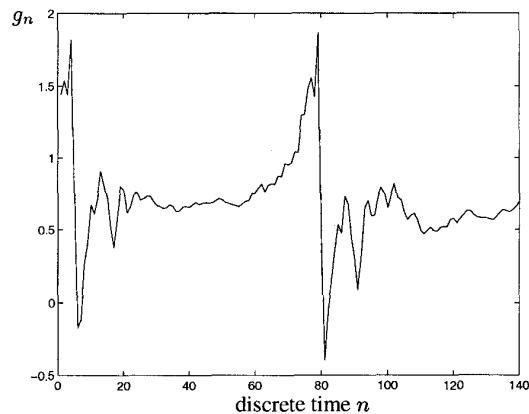


Figure 4: Glottal Signal Estimate.

3. EXAMPLE APPLICATION

The system that will be used here to demonstrate the proposed approach utilizes the least-squares (LS) procedure which is known to converge after about $2p$ data samples [3]. Consequently, it is capable of extracting accurate estimates from much shorter closed glottis regions (e.g., from high-pitched speakers) than is possible with the Covariance method used in glottal inverse filtering. The forgetting factor is set to 0.97 for all analyses. Values much larger would not provide fast enough convergence resulting in an uninteresting glottal signal with little deviation from unity. Values much smaller produced a noisy glottal signal. Since the proposed approach assumes that the speech input is voiced, a preprocessing routine is used that determines if the current frame consists of unvoiced speech. If so, standard LPC analysis is performed to obtain a prediction vector for these frames, and the glottal signal computation is bypassed altogether.

3.1. Examples of The Glottal Signal

This section describes the glottal behavior as was observed

in the time-varying frequency response in Figure 1, but now as it affects the glottal signal $\{g_n\}$ derived using the proposed system. Figures 4 and 5 show glottal signals extracted from a subset of the diverse database described in the appendix as they say "o" in the word "home". Speaker GLB is an adult female and BSB and JTA are male children.

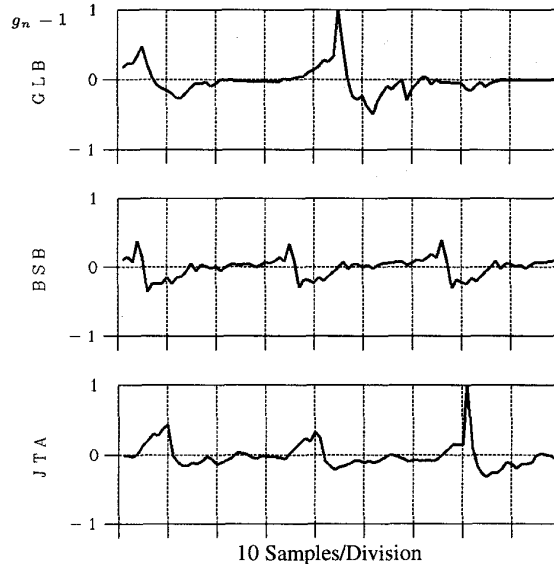


Figure 5: Glottal Signals From The Vowel in "Home".

The most distinctive event in the computed glottal signals is the open glottis interval which contains the sharp peak used by the proposed algorithm to identify each pitch period. The open glottis interval ends when glottal excitation occurs. In the glottal signals, this event produces the sharp negative slope at the end of the open glottis interval. The relatively flat interval that follows and extends to the next glottal opening phase is the closed glottis interval. The fact that the value of the original glottal signal (prior to changing the bias by negative one) during this interval is approximately unity implies that the formants are quite stationary and similar to a_{n_c} in Section 2.1.

Upon observation of the glottal signals in the various plots, one can see that these general characteristics are quite universal for this yet diverse population of speakers. Perhaps the most obvious difference between the plots is the fundamental frequency of the signals due to the large pitch range of the database of speakers.

Another noticeable difference between these speakers is the sharpness of the valley at the instant of glottal closure. Obviously this valley is much more distinguishable for the glottal signal for WES (in Figure 4) than it is for JTA (Figure 5). The glottal opening pulse size can vary between speakers as well. For example the pulse for WES is at least 15 samples, compared to less than 10 samples for GLB. One should notice, however, that this increase in opening pulse duration is not proportional to the increase in pitch period between the speakers. The measurement used to describe this characteristic is referred to as the open quotient (OQ), and is formally defined as the open phase duration divided by the pitch period. So in other words, although the opening pulse duration is smaller for GLB, the open quotient is at least as large as for WES.

The characteristics discussed above correspond to findings in

other studies. The lack of sharpness at glottal closure, as was observed for speaker JTA, was found to be characteristic of breathy voices [2]. Studies also confirm that the open quotient is generally larger for female than for male speakers [2, 6].

3.2. A Simple SIV Example

A simple speaker identity verification (SIV) approach was used to evaluate the proposed analysis system and the effectiveness of the features it extracts at characterizing or distinguishing speakers. The database used in the test was collected from a diverse group of male and female adults and children (refer to the appendix), and provided a wide variation of pitch. The SIV approach assumed vocabulary dependence so that, given either the vocal tract or vocal cord features, two different recordings of the same word are always compared.

For the case of the vocal tract model, two recordings of a word were compared by converting the prediction coefficients to cepstral coefficients for use in the Euclidean distance formula with inverse variance weighting. Due to availability, dynamic time-warping was used for aligning the features. A justification for choosing this matching technique is that it has been found to yield similar SIV accuracy as hidden Markov modeling approaches for short verification utterances when training data is limited [7].

For the case of the vocal cord model, the proposed approach extracts segments of the glottal signal (each of which contains at least one pitch period) and chooses one determined to be the best representative of the glottal signal based on waveform comparisons between each segment. These time-domain comparisons also used the inverse variance weighting for the Euclidean distance. The score between two glottal signal segments was determined by selecting the smallest distance while the alignment of the signals was varied a sample at a time up to a quarter of the period length.

The proposed approach achieved an equal-error rate (EER) of 20%. In comparison, the same experiment was performed using a conventional linear predictive analysis algorithm (as described in [8]) with 30 ms non-overlapping frames as well, which resulted in an EER of 18%. In order to evaluate the quality of the speaker information contained in the proposed glottal signal, the EER was computed for each speaker using all features, and again using only the linear prediction vocal tract features. For 21 of the 29 speakers, it was found that the glottal signal improved the EER by as much as 2.28%, but degraded it by close to 14%, 8% and 5% for three of the remaining eight speakers. The difficulty encountered with all three speakers (an adult male, female child and another adult male, respectively) originated in the glottal signal extraction and v/uv detection routines as a result of weak voicing. Hence the glottal signal contains useful information for some speakers, but not for others. This case for speakers with breathy characteristics is an obvious one where it does not make sense to attempt to extract and use any glottal information because of the risk of noisy parameters.

4. SUMMARY

A procedure has been proposed that takes advantage of the fast convergence properties of adaptive filtering in order to model the nonstationarities due to the vocal tract and rapidly time-varying vocal cord behavior. The vocal tract is modeled by extracting a single vector of autoregressive coefficients per frame. Unlike in LPC analysis, it is chosen from within the closed glottis interval. To model vocal cord behavior, the proposed system extracts a glottal signal which is obtained by initially computing the unit

step response through the time-varying filter. The glottal signal is then derived by passing the unit step response through the time-invariant inverse filter defined by the vocal tract coefficients for the respective frame. Glottal signals extracted by the proposed procedure have been analyzed for several speakers. The effectiveness of this vocal cord information at distinguishing speakers was evaluated using a simple SIV approach, and was found to yield reasonable performance, comparable to standard autocorrelation analysis.

5. APPENDIX

The speech database used in this paper was collected from a church congregation in Camarillo, California, which provided a diverse group of speakers consisting of 13 boys from 4 to 16, 5 girls from 7 to 17, 7 men from 38 to 82, and 4 women from 28 to 75 years of age. Many of these subjects were born and raised in various regions across the United States.

The vocabulary was a subset of the phonetically diverse words presented by Velius [8]. These words were recited by a given subject once per recording session. Subjects took part in ten sessions which occurred over a six month period.

Recordings were performed in a quiet room onto a 16 bit sound card at a sampling rate of 22.05 kHz. The recordings were then scaled and decimated to 8 kHz using three multirate FIR filter stages. Endpoint detection was then performed on the decimated speech, removing most of the nonspeech sections of the records. The parameters of this endpoint detection routine (documented in [1]) were set conservatively to ensure that none of the spoken words were removed.

6. REFERENCES

- [1] B. L. Berg and T. C. Feustel, "A complete endpoint detection routine for use in speaker identity verification", Technical Memo. TM-ST5-021674, Piscataway, NJ, Bellcore, 1992.
- [2] D. G. Childers and C. K. Lee, "Voice quality factors: analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, Vol. 90, pp. 2394-2410, Nov. 1991.
- [3] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 1991.
- [4] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.
- [5] M. J. Hunt, J. S. Bridle, and J. N. Holmes, "Interactive digital inverse filtering and its relation to linear prediction methods," *Proc. IEEE ICASSP-78*, pp. 15-18, 1978.
- [6] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, Vol. 87, pp. 820-857, Feb. 1990.
- [7] A. E. Rosenberg, C. Lee, and S. Gokcen, "Connected word talker verification using whole word hidden markov models," *Proc. IEEE ICASSP-91*, pp. 381-384, 1991.
- [8] G. Velius, "Variants of cepstrum based speaker identity verification," *Proc. IEEE ICASSP-88*, pp. 583-586, 1988.
- [9] D. Y. Wong, J. D. Markel and A. H. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 27, pp. 350-355, Aug. 1979.