THE DYPSA ALGORITHM FOR ESTIMATION OF GLOTTAL CLOSURE INSTANTS IN VOICED SPEECH

Anastasis Kounoudes, Patrick A. Naylor, Mike Brookes

Department of Electrical and Electronic Engineering, Imperial College, London, UK.

ABSTRACT

We present the DYPSA algorithm for automatic and reliable estimation of glottal closure instants (GCIs) in voiced speech. Reliable GCI estimation is essential for closed-phase speech analysis, from which can be derived features of the vocal tract and, separately, the voice source. It has been shown that such features can be used with significant advantages in applications such as speaker recognition. DYPSA is automatic and operates using the speech signal alone without the need for an EGG or Laryngograph signal. It incorporates a new technique for estimating GCI candidates and employs dynamic programming to select the most likely candidates according to a defined cost function. We review and evaluate three existing methods and compare our new algorithm to them. Results for DYPSA show GCI detection accuracy to within ±0.25ms on 87% of the test database and fewer than 1% false alarms and misses.

1. INTRODUCTION

Conventional speech analysis procedures use autoregressive modeling in LPC-based approaches or spectral/cepstral estimation in transform-based approaches. These methods work well for cases where only a spectral estimate is required but do not explicitly deconvolve the transfer function of the vocal tract, which is assumed quasi-stationary, from the excitation signal, which can be modeled as a quasi-periodic signal in voiced speech and a noise-like excitation in unvoiced speech. Consequently, the features extracted by conventional analysis methods represent the combined effects of source and tract. However, in several important applications of speech processing, including speaker recognition and speech coding, it is advantageous to extract reliable estimates of the vocal tract transfer function and, separately, the properties of the voice source. The algorithm described here for GCI estimation provides the segmentation of the larynx cycle necessary for the solution to this latter estimation problem.

It has been shown [1] that the voice source signal can be deconvolved from the speech signal using multicycle closed-phase inverse filtering (MCIF) [2] and that the resulting signal can be successfully parameterized and used, for example, to provide additional features in text-dependent speaker verification. However, this blind deconvolution relies on accurate segmentation of the voiced speech larynx cycle into closed and open phases of the glottis. To date, this has required the use of contemporaneous laryngographic recordings [3, 4] (EGG) from which to derive glottal closure instants. Because the EGG signal is not normally available in practical applications, there exists a strong motivation to develop techniques for extracting GCIs from the speech signal alone. In this paper we present such a technique for estimating GCIs, known as the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA), that enables the use of voice-source features and accurate vocal tract transfer function estimates in the domain of practical applications.

2. SEGMENTATION OF THE LARYNX CYCLE

Several algorithms have been proposed for determining glottal closure instants from a speech waveform. One of the earliest approaches [5] derived GCIs from the autocovariance matrix of the speech signal and later work [6] used the minimum energy in the LPC residual. As a development of [5], the GCIs in [7] are identified as the maxima of the Frobenius Norm of the signal matrix. The authors reported significant improvements in performance, computational complexity and noise robustness. The method proposed in [8] estimates the location of the excitation within an analysis frame as the average value of the group delay. Recently, work on energy flow in the lossless-tube model has been reported [10] and it was suggested that the signal representing acoustic input power at the glottis can be used to determine the instants of glottal closure and opening.

In this paper, the APLAWD database [11] has been used to perform comparative evaluations of three methods for estimating GCIs in voiced speech using Wong's LPC residual (LPCR) [6], the Frobenius Norm (FN) [7] and the Group Delay (GD) [8]. APLAWD contains phonetically balanced speech from 5 male and 5 female talkers as well as EGG recordings from which, after time-alignment, reference GCIs have been extracted using the HQTx algorithm [9]. The alignment of estimated GCIs to the reference GCIs used dynamic programming to minimize total absolute detection error. For each method

0-7803-7402-9/02/$17.00 ©2002 IEEE

we measure false alarm rate (FAR) and miss rate (MR). We also assess accuracy of detection by first computing the distribution $\zeta = \left(t_d(i) - t_{ref}(i)\right)$ where $t_d(i)$ and $t_{ref}(i)$ are the instants of the detected and reference GCI in cycle i respectively, excluding misses and false alarms. The mean value of ζ represents a constant time offset that can be easily taken into account and subsequently corrected. We define accuracy as the percentage of the distribution that lies within $\pm T_b$ of the mean. We have chosen to study accuracy to $T_b = 0.25\text{ms}$ as this corresponds to the practical limit on accuracy required for reliable closed-phase LPC analysis [12].

Figure 1 shows distributions, ζ, measured in ms for each of the methods studied and aggregated over all talkers. Tables 1 and 2 shows comparative results for FAR, MR and accuracy of the methods.

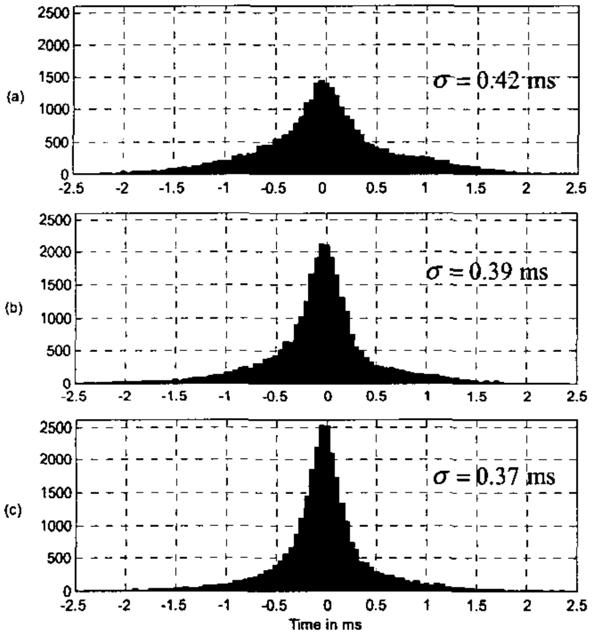

Fig. 1. Distribution ζ of measured detection accuracy for (a) Wong's LPC residual method, (b) the Frobenius Norm method and (c) the Group Delay method.

We observe from Fig. 1 and Table 1 that the GD method outperforms the other methods in our tests. Its accuracy is relatively good (75%) but it nevertheless exhibits a high rate of false alarms and misses (>10%). From this position we are motivated to improve FAR and MR, as described in the following Section, so that the good accuracy in the results of the GD method can be exploited reliably.

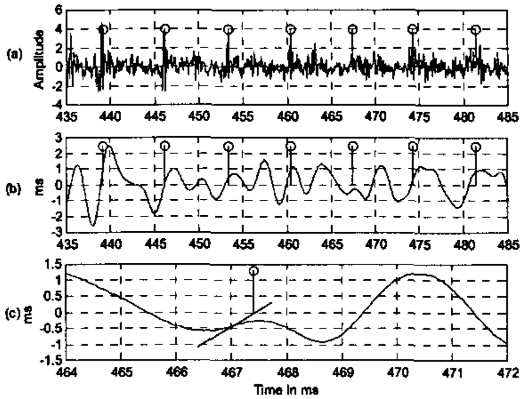

Fig. 2. The enhanced GD algorithm: (a) LPC prediction residual and reference GCIs from EGG, (b) the phase-slope function and reference GCIs from EGG (c) detail showing the projection of "missed" zero-crossings.

3. DYPSA – A DYNAMIC PROGRAMMING APPROACH TO GCI SELECTION

DYPSA uses dynamic programming (DP) [13] to determine the most likely combination of GCIs from a set of GCI candidates. The candidates are obtained using a new technique that operates on the phase-slope function of the GD algorithm [8], which is defined as the slope of the unwrapped phase of the short-time Fourier transform of the linear prediction residual. In [8], instants of glottal closure are identified as positive-going zero-crossings in this phase-slope function. In this work, we have identified and included additional candidates, the omission of which would otherwise cause GCI misses arising from maxima or minima that fail to cross zero. Whenever a minimum is followed by a maximum without an intervening zero-crossing, the midpoint between the two extrema is identified and its position projected with unit slope onto the time axis, under the assumption that the ideal phase-slope at a zero-crossing is unity [8]. In this way, by defining the set of GCI candidates to be the union of all positive going zero-crossings and projected zero-crossings, the number of detection misses has been significantly reduced. This procedure is illustrated in Fig. 2.

The problem of GCI estimation is now considered as a minimization of a cost function using DP. To reduce the storage and computation requirements of the algorithm, we use a search strategy in which only the N-best path segments are retained at each stage of the DP. A reasonable trade-off between complexity and performance was found when N=3.

The factors used in the construction of the cost function are based on the attributes of the GD and FN methods as well as the periodic behavior of the vocal folds. The cost function is defined as

$$C_{p,q,r} = w_{pitch} C_{pitch_{p,q,r}} + (w_{FN} - A_d) C_{FN_{p,r}} + (w_h - A_d) C_h + (w_a + A_d) C_{a_{p,r}} + ZCB$$

where r, q and p represent respectively the current and previous two GCI candidates. We have empirically determined the weights in the cost function as $[w_{pitch}, w_{FN}, w_h, w_a] = [0.5, 0.5, 0.5, 0.25]$. Additionally, we adjust the weights for voiced and unvoiced speech using the adaptation factor,

$$A_d = \begin{cases} 0, & \text{voiced} \\ 0.25, & \text{unvoiced} \end{cases}$$

for which we have employed a voiced/unvoiced detector based on the ratio of short-time speech energy to zero-crossing rate [14]. The *ZCB* term adds a penalty when the GCI candidate arises from the projection of a turning point onto the time-axis as described in Section 3.

$$ZCB = \begin{cases} 0, & \text{current candidate is a positive zero-crossing} \\ 0.2, & \text{current candidate is a projection} \end{cases}$$

The individual components of the cost function are defined as follows.

Pitch Deviation Cost is a function of adjacent GCI candidates taken from the last three stages of the path segment under consideration and is defined as

$$C_{pitch_{p,q,r}} = -\chi_{1,0.2}\left(\frac{\min\left[(t_r - t_q), (t_q - t_p)\right]}{\max\left[(t_r - t_q), (t_q - t_p)\right]}\right) + 0.5$$

where $\chi_{\mu,\sigma} = e^{-(y-\mu)^2/2\sigma^2}$ and

t_j is the time of occurrence of candidate GCI(j).

This cost increases with pitch deviation between successive larynx cycles as shown in Fig. 3 and is based on the assumption of smooth variation in pitch over short segments of voiced speech.

Frobenius Norm Amplitude Consistency Cost is formulated as

$$C_{FN_{q,r}} = 0.5 - \frac{\min(FN_q, FN_r)}{\max(FN_q, FN_r)}$$

where FN_j is the Frobenius Norm [7] of the speech data matrix value estimated over a 3ms window centered on GCI candidate j. This cost increases with variation in *FN* between successive cycles. Since the *FN* of the speech data matrix is normally much larger at instants of true excitation than at false alarms, candidate GCIs for which the *FN* is significantly different, usually smaller, than neighboring GCIs are more likely to be false alarms and are penalized accordingly.

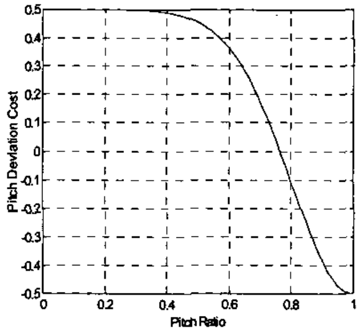

Fig. 3. Pitch Deviation Cost as a function of pitch ratio between successive cycles.

Ideal Phase Slope Function Deviation Cost provides an indication of the "goodness" of the phase-slope function [8] and is defined as

$$C_{h_{p,r}} = \left(0.5 - \frac{\min(\gamma_p, \gamma_r)}{\max(\gamma_p, \gamma_r)}\right).$$

Our experiments have shown that it is quite common to obtain two positive going zero-crossings per cycle in the phase-slope function. It is interesting to consider, as further work, whether the first corresponds to the glottal closure and the second to opening. We define the *ideal phase-slope* at a positive zero-crossing to be a straight line of unit gradient, which corresponds to an impulsive excitation at the GCI. Under this definition, we have found that GCI candidates corresponding to true GCIs have a phase-slope function significantly closer to the ideal than other candidates. We therefore formulate the phase-slope deviation cost by computing the sum square error, γ, between the measured phase-slope function and the ideal phase-slope, calculated over a short (0.85ms) window centered on the zero-crossing.

Speech Waveform Similarity Cost uses the normalized cross-correlation, $NCorr_{p,r}$, estimated using 10ms speech segments centered at the GCI candidates p and r

$$C_{a_{p,r}} = -NCorr_{p,r} / 2.$$

During voicing, it is common to find that the speech waveform near an instant of excitation is well correlated to the waveform at the previous excitation. We therefore apply a high cost to any candidate GCIs that occur when the speech signal is significantly uncorrelated with the signal at the previous GCI. This serves effectively to penalize any candidates that occur, for example, part way through a larynx cycle.

4. SIMULATION RESULTS

The tests described in Section 2 using the APLAWD database have been repeated on the DYPSA algorithm. Fig. 4 shows the corresponding results. Tables 1 and 2 show comparisons of the performance criteria of FAR, MR and accuracy, defined as the percentage of GCIs detected within 0.25 ms of the reference.

We have additionally tested the proposed method in the presence of noise and find that, for an SNR of 30dB, the accuracy drops by about 5% and the FAR and MR both increase to around 2%. This shows substantially more robustness than the existing methods that degrade badly in noise [8].

Performance (all phonemes)		LPCR (%)	FN (%)	GD (%)	DYPSA (%)
FAR	Males	28.2	25.1	10.7	0.9
	Females	42.7	21.6	9.1	0.8
MR	Males	1.0	0.9	1.5	0.1
	Females	7.4	3.8	3.8	0.1
Percent Accuracy ±0.25 ms	Males	57.3	69.4	77.0	90.4
	Females	32.3	52.0	73.2	85.1

Table 1. Performance averaged across all phonemes.

Accuracy to ±0.25 ms (all talkers)	LPCR (%)	FN (%)	GD (%)	DYPSA (%)
/a/	47.3	62.2	73.4	91.5
/e/	40.8	58.8	69.7	82.1
/i/	47.9	65.7	77.6	90.7
/o/	42.3	58.7	77.5	89.3
/u/	35.1	51.1	70.8	83.6

Table 2. Accuracy as percentage of GCIs detected within 0.25ms of the reference, averaged across all talkers.

5. DISCUSSION AND CONCLUSIONS

The DYPSA algorithm for extracting instants of voiced speech excitation, GCIs, is based on an enhancement of the GD algorithm and uses DP to minimize a cost function so as to eliminate almost all false alarms and misses in the detection. It has been shown that the new method gives significantly better overall accuracy. The FAR obtained is typically less than 1% and MR less than 0.1% compared with around 10% and 2% respectively for the original GD method. Accuracy of the new method is 80-90% depending on talker and phoneme compared to 70-80% for the GD method. Using segmentation based on DYPSA, features of the voice source can now be extracted and used to good effect in high quality speech analysis, speaker recognition and other related applications.

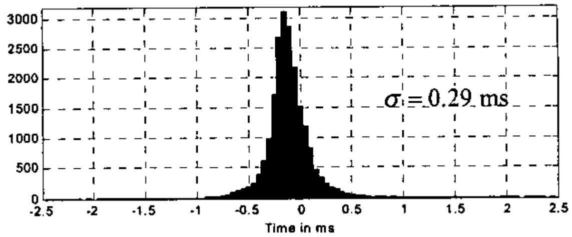

Fig. 4. Distribution ζ of measured detection accuracy for the proposed algorithm.

6. REFERENCES

[1] A. Neocleous and P.A. Naylor, "Voice Source Parameters for Speaker Verification," in *Proc. European Signal Processing Conf.*, pp. 697-700, 1998.

[2] D.M. Brookes, D.S. Chan, "Speaker Characteristics from a Glottal Airflow Model using Glottal Inverse Filtering," *Proc. Institute of Acoustics*, vol. 15, pp. 501-508, 1994.

[3] E.R.M. Abberton, D.M. Howard and A.J. Fourcin, "Laryngographic assessment of normal voice: a tutorial", Clinical Linguistics and Phonetics, pp. 281-296, 1989.

[4] A.K. Krishnamurthy, D.G. Childers, "Two-channel speech analysis", *IEEE Trans. ASSP.*, vol. 34, pp. 730-743, 1986.

[5] H.W. Strube, "Determination of the Instant of Glottal Closures from the Speech Wave," *J. Acoust. Soc. Am.*, vol. 56, pp. 1625-1629, 1974.

[6] D.Y. Wong, J.D. Markel and A.H. Gray, "Least Squares Inverse Filtering from the Acoustic Speech Waveform," *IEEE Trans. ASSP.*, vol. ASSP-27, pp. 350-355, 1979.

[7] C.X. Ma, Y. Kamp and L.F. Willems, "A Frobenius Norm Approach to Glottal Closure Detection from the Speech Signal," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 258-265, 1994.

[8] R. Smits and B. Yegnanarayana, "Determination of Instants of Significant Excitation in Speech Using Group Delay Function," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 325 -333, 1995.

[9] M. A. Huckvale, "Speech Filing System," http://www.phon.ucl.ac.uk/resource/sfs/.

[10] D.M. Brookes and H.P. Loke, "Modelling Energy Flow in the Vocal Tract with Applications to Glottal Closure and Opening Detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 213-216, 1999.

[11] G. Lindsey, A. Breen, S. Nevard, "SPAR's Archivable Actual-word Databases," Department of Phonetics and Linguistics, University College, London, June 1987.

[12] A. Neocleous, "Speaker Verification using Voice Source Parameters", PhD Thesis, Imperial College, 2000.

[13] H. Ney, "A Dynamic Programming Technique for Nonlinear Smoothing," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 62-65, 1981.

[14] A. Kounoudes, "Epoch Estimation for Closed-Phase Analysis of Speech," PhD Thesis, Imperial College, 2001.