Speech Classification of Glottal Closure Instants in the Dypsa Algorithm

Hania Maqsood*, Patrick A. Naylor**
* Bahria Institute of Management and Computer Sciences, Islamabad, Pakistan
haniamaqsood@yahoo.com
**Imperial College, Department of Electrical and Electronic Engineering, London, UK
p.naylor@imperial.ac.uk

Abstract

The Dynamic Programming Projected-Phase Slope Algorithm (DYPSA) detects glottal closure instants (GCI) in speech signals. We present an improvement in the algorithm in which a voiced/unvoiced/silence discrimination measure is applied in order to reduce the spurious GCIs detected incorrectly for silence and unvoiced speech. Speech classification is addressed by formulating a decision rule for the glottal closure instant candidates which classifies the candidates as voiced or non-voiced (the ones occurring in silence and unvoiced speech) on the basis of feature measurements extracted from the speech signal alone. The technique of Dynamic Programming is then employed in order to select an optimum set of epochs from the GCI candidates. The algorithm has been tested on the APLAWD speech database with 87.23% improvement achieved in reduction of spurious GCIs.

Introduction

The classical model of human speech production system is generally represented by a linear source tract model excited by a quasi-periodic signal or a noise like waveform. In several important applications of speech processing, it is advantageous to work with the vocal tract and the excitation signal independently. The separation of the vocal tract from the source excitation is seen as a distinguishing problem between opening and closing of the glottis (V-shaped opening in the vocal cords). Each cycle of voiced speech can be divided into a close phase, during which air flow through the glottis is blocked by closure of the vocal folds, and an open phase during which the vocal folds are open. The main acoustic excitations of the vocal tract occur at the glottal closure instants (GCIs). Having the ability to identify the instants of glottal closure enables the use of larynx synchronous processing techniques such as closed-phase linear predictive coding (LPC) analysis [1] and closed-phase glottal inverse filtering [2]. These techniques make it possible to separate the characteristics of the glottal excitation waveform from those of the vocal tract filter and to treat the two independently in subsequent processing. Applications include low bit-rate coding [3] [4], data-driven techniques for speech synthesis [5], prosody extraction [6], speaker normalization and speaker recognition. The DYPSA algorithm is a recently proposed technique for identifying GCIs and will be discussed in the following section. In this paper, we describe a modified version of the algorithm which maintains all the advantages of DYPSA's high accuracy in voiced speech but overcomes the problem of erroneously detected spurious GCIs for silence and unvoiced speech encountered in the current form of the algorithm. The approach will involve defining 3 classes of speech as voiced, unvoiced and silence. In practical applications, true silence is always disturbed by the presence of noise. Therefore, we use the term 'silence' in this paper to mean the absence of speech, such as occurs outside speech endpoints or during short pauses. We shall be using the term non-voiced to cater for both silence and unvoiced speech.

Review of the DYPSA algorithm

The Dynamic Programming Projected-Phase Slope Algorithm (DYPSA) is an automatic technique for estimating GCIs in voiced speech from the speech signal alone [7]. DYPSA involves the extraction of candidate GCIs using phase-slope function as presented in [8]. The GCIs are identified from this phase-slope function as positive-going zero-crossings (PZC). DYPSA also involves identification of additional candidates, which may have been missed if the phase-slope function fails to cross zero appropriately. An optimum set of epochs (candidates) is then selected by minimizing a cost function using N-best Dynamic Programming (DP) technique as presented in [9] [10]. The cost function comprises of speech waveform similarity cost, pitch deviation cost, projected candidate cost, normalized energy cost and the ideal phase-slope function deviation cost.

The accuracy of DYPSA has been tested on the Archivable Priority List Actual Word Database (APLAWD) [11] with the reference GCIs extracted from the EGG (electroglottograph) signal using the HQTx program from the Speech Filing System software suite [21]. APLAWD is an English speech database consisting of 146 one- or two-word items and five sentences. The evaluations have been carried out using only the sentence subset of APLAWD recorded at a sample rate of 20 KHz. The database includes ten repetitions from each of ten British English speakers (five male, five female) of the following sentences for a total of 500 utterances. The sentences are: "George made the girl measure a good blue vase"; "Why are you early you owl?"; "Cathy hears a

Corresponding Author: Hania Maqsood

NUST Publishing, © (2008), ISSN: 2070-9900

voice amongst SPAR's data"; "Be sure to fetch a file and send their's off to Hove"; "Six plus three equals nine".

A comparative evaluation of DYPSA with the previous techniques such as [12], [13] and [8], has shown significantly enhanced performance with identification of 95.7% of true GCIs in voiced speech.

However DYPSA, in its current form is not able to distinguish when voiced speech is present and the algorithm detects spurious GCIs for non-voiced speech. For DYPSA to operate independently over speech segments containing both voiced and non-voiced speech, we need to detect the regions of voicing activity. This is viewed as a classification problem between voiced and non-voiced speech. The solution involves incorporating a voicing decision for the GCI candidates within the algorithm. The GCI candidates identified as occurring in the unvoiced speech segments are then removed.

Identification of GCI candidates

The speech signal with sampling frequency 20kHz is passed through a 1st order pre-emphasis filter with a 50 Hz cut-off frequency to flatten the speech spectrum. The speech signal is then processed using autocorrelation Linear Predictive Coding (LPC) [3] of order 22 with a 20 ms Hamming window overlapped by 50%. The pre-emphasized speech is inverse filtered with linear interpolation of the LPC coefficients for 2.5 ms on either side of the frame boundary. Given the residual signal u(n), and applying a sliding M-sample Hamming window w(m), as defined in [7], we obtain frames of data in vicinity of each sample n as:

$$x_n(m) = \begin{cases} w(m)u(m+n), & m = 0,....,M-1 \\ 0, & otherwise \end{cases} \quad (1)$$

with Fourier transform $X_n(\omega)$. The phase slope function was defined in [8] to be the average slope of the unwrapped phase spectrum of the short time Fourier transform of the linear prediction residual. The phase slope function defined in [7] is:

$$\tilde{\tau}_n(\omega) = \frac{d\arg(X_n(\omega))}{d\omega} \quad (2)$$

DYPSA identifies the instants of glottal closure as the positive-going zero-crossings of the phase slope function. In studying the phase slope function it is observed that GCI events can go undetected because the phase slope function fails to cross zero appropriately, even though the turning points and general form of the waveform are consistent with the presence of an impulsive event indicating a GCI. To recover such otherwise undetected GCI candidates, DYPSA relies on a phase-slope projection technique. In this method, whenever a local minimum is followed by a local maximum without an interleaving zero-crossing, the mid point between the two extrema is identified and its position is projected with unit slope onto the time axis. This technique is presented in [7] and draws on the assumption that, in the absence of noise the phase slope at a zero-crossing is unity. The final set of GCI candidates is defined as a union of all positive-going zero-crossings and the projected zero-crossings.

Dynamic programming

The selection of true GCIs from set of voiced candidates is performed by minimizing a cost function using N-best dynamic programming [9][10]. The procedures maintain information about N most likely hypothesis at each step of the algorithm. The value of N is chosen as 3 following the approach in [7]. Cost function to be minimized is defined as:

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \lambda^T c_{\Omega}(r) \quad (3)$$

where Ω is a subset of GCIs selected from all GCI candidates, $|\Omega|$ is the size of Ω, r indexes the elements of Ω, $[.]^T$ represents transpose and λ is a vector of weighting factors defined from [7] as:

$$\lambda = [\lambda_A, \lambda_P, \lambda_J, \lambda_F, \lambda_S]^T = [0.8,0.5,0.4,0.3,0.1]^T \quad (4)$$

The elements of the cost vector evaluated for the rth GCI are:

$$c_{\Omega}(r) = [c_A(r), c_P(r), c_J(r), c_F(r), c_s(r)]^T \quad (5)$$

where $c_A(r)$ represents the speech waveform similarity cost, $c_P(r)$ represents the pitch deviation cost, $c_J(r)$ represents the projected candidate cost, $c_F(r)$ represents the normalized energy cost and $c_S(r)$ represents the ideal phase-slope function cost. The elements of the cost function all lie in the range [-0.5, 0.5] and a low cost indicates a true GCI. The advantage of using the DP cost function is that it effectively penalizes GCI candidates in a way that in most cases all but one candidate per larynx cycles is rejected. For further details the reader is referred to [7].

Voiced, unvoiced, silence classification

Segments of speech can be broadly classified into three main classes: silence, unvoiced and voiced speech. Silence is the part of signal where no speech is present and is generally encountered in the beginning or ending of speech recordings and between talk spurts. Unvoiced sounds result when the excitation is a noise-like turbulence produced by forcing air at high velocities through a constriction in the vocal tract while the glottis is held open. On the other hand voiced sound is produced by the vocal fold periodic vibration. The technique adopted for speech classification takes into consideration the statistical distributions and characteristics features of the three classes. The main components of the classifier as represented by Fig. 1 are

(1) feature extraction, (2) Gaussian mixture modeling and (3) the decision algorithm.

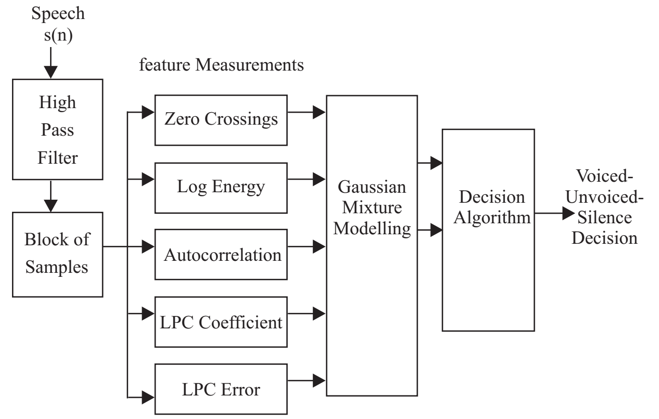

Fig. 1. Block diagram of voiced-unvoiced-silence detector

Feature extraction

Prior to analysis, the speech signal is high-pass filtered at approximately 200 Hz to remove any dc or noise components. Frames of 10 ms duration are then defined centered on each GCI candidate found for DYPSA as described in section 2.1. For every frame a set of features are extracted. The choice of the features set is based on experimental evidence of variations between classes and from the knowledge of human speech production model. The five features used in implementing the classifier, based on [14] are:

1 Zero-Crossing Rate: Voiced speech usually shows a low zero-crossing rate while unvoiced speech has a concentration of energy at high frequencies and typically exhibits a high zero-crossing rate. The zero-crossing count for silence varies from one speaking environment to another based on the background noise.

2 Log Energy is defined as:

$$E_s = 10 * \log_{10}\left(\varepsilon + \frac{1}{N}\sum_{n=1}^{N} s^2(n)\right) \tag{6}$$

where ε = 10^{-5}, is a small positive constant added to prevent computing log of zero. The energy of voiced sounds is much higher than the energy of silence. The energy of unvoiced sounds is usually lower than for voiced sounds, but often higher than for silence.

3. Normalized Autocorrelation Coefficient is defined as:

$$C_1 = \frac{\sum\limits_{n=1}^{N} s(n)s(n-1)}{\sqrt{\left(\sum\limits_{n=1}^{N} s^2(n)\right)\left(\sum\limits_{n=0}^{N-1} s^2(n)\right)}} \tag{7}$$

Parameter C1 correlates adjacent speech samples and varies between - 1 and +1. As adjacent samples of voiced speech waveform are highly correlated therefore C1 is close to unity. On the other hand, the correlation is close to zero for unvoiced speech.

4. First Predictor Coefficient from Linear Predictive Analysis: It was shown by Atal [14] that the first predictor coefficient obtained from LPC, is identical (with a negative sign) to the cepstrum of the log spectrum at unit sample delay. Since spectra of the three classes-voiced, unvoiced and silence differ considerably, so does the first LPC coefficient and thus the first predictor coefficient is used as a discrimination measure between the three classes.

5. Normalized Prediction Error: A by-product of the Linear Predictive analysis is the prediction error signal defined (in dB) [15] as:

$$E_p = E_s - 10 * \log_{10} \left(10^{-6} + \left| \sum_{k=1}^{p} a_k \phi(0,k) + \phi(0,0) \right| \right) \tag{8}$$

where E_s is the log energy defined in (6) and $\phi(i,k) = \frac{1}{N}\sum_{n=1}^{N} s(n-i)s(n-k)$ is the (i, k) term of the covariance matrix of speech samples. Round off errors may yield a small negative value and 10^{-6} is added to prevent computation of log of a negative number. The normalized prediction error is considered as a measure of the uniformity of the spectrum. The spectrum of voiced speech has a well-defined formant structure which results in higher prediction error as compared to unvoiced speech or silence.

Out of the five parameters discussed above, none are sufficiently reliable to give robust classification in the face of noise, speaker variation, speaking style and so forth as confirmed by earlier studies [16]. Therefore our decision algorithm makes use of all five features to optimally combine their contributions in differentiating between the three classes.

Gaussian mixture modelling

It is assumed that the features for each class are from a multidimensional Gaussian distribution where each class is modeled as a Gaussian-shaped cluster of points in feature space (in our case, 5-dimensional space). This assumption has the advantages of computational simplicity as the decision rule is determined by the mean vector μ and covariance matrix C estimated from the feature vector itself. In order to estimate the parameter set we employ the K-mean clustering algorithm followed by iterations via Expectation Maximization (EM) algorithm proposed by Dempster [17]. The K-mean Algorithm partitions the points of a data matrix into K clusters by minimizing the distance to the nearest cluster centre (centroid). This process is repeated till the cluster centers converge [18] [19]. The EM Algorithm then maximizes the log likelihood from incomplete data in order to estimate the parameters of the distribution [20]. For simplification of computation the individual clusters are not represented with full covariance matrices, but only the diagonal approximations. Our experiments have shown that no significant improvement is obtained from using full covariance matrices in this context.

Decision algorithm

We assume that the joint probability density function of the possible values of the measurements for the ith class is a multidimensional Gaussian distribution, where i= 1, 2, 3 corresponds to the voiced, unvoiced and silence classes respectively. Let x be a d-dimensional column vector (in our case, d=5) representing the measurements. Then the d-dimensional Gaussian density function for x with mean vector μ_i and covariance matrix C_i is given by:

$$g_i(x) = (2\pi)^{-d/2} \mid C_i \mid^{-1/2} \exp[-\frac{1}{2}(x-\mu_i)^T C_i^{-1}(x-\mu_i)] \quad (9)$$

where C_i^{-1} is the inverse of the matrix C_i, $\mid C_i \mid$ is the determinant of C_i. We define the normalized voicing measure as:

$$\Psi_{vus} = \frac{g_1(x)}{g_1(x)+g_2(x)+g_3(x)} \quad (10)$$

From the definition in (10), the GCI candidates occurring in the voiced segments of speech get assigned a higher score. To simplify computation, taking the natural log on both sides of (9) we obtain

$$\ln(g_i(x)) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\left|C_i\right| - [-\frac{1}{2}(x-\mu_i)^T C_i^{-1}(x-\mu_i)] \quad (11)$$

We define:

$$\ln(\Psi_{vus}) = \ln(g_1(x)) - \ln(g_1(x) + g_2(x) + g_3(x)) \quad (12)$$

The candidates in the voiced regions are assigned a high score whereas for the unvoiced speech and silence we obtain a low score (close to zero). The question now remains as to the choice of a threshold value for the voicing score. The threshold of 0.1 has been chosen by empirically as suitable for the APLAWD database. GCI candidates with scores below this threshold are excluded from further processing. This avoids DYPSA from giving spurious GCIs during unvoiced speech or silence and also simplifies the computation required for the DP routine within DYPSA.

Experiments and results

For the performance evaluation of DYPSA we require reference GCIs which are obtained from the EGG signal. The speech and the EGG signal are first time-aligned and reference GCIs are then extracted from the EGG signal using HQTx algorithm [21]. The performance comparison is carried out based on the HQTx markers (indicating 'ground truth' GCIs in the speech waveform) and the GCIs obtained from DYPSA by the dynamic programming technique.

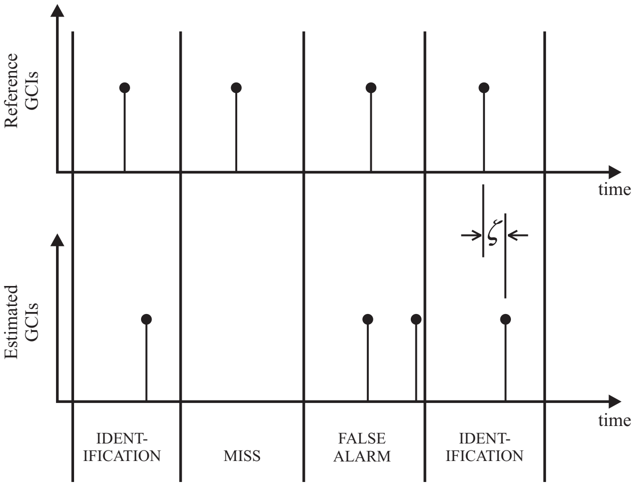

Fig. 2. Definition of evaluation metrics. The dotted lines depict a frame defined around each reference GCI marker to indicate a larynx cycle (after [7])

In order to access the performance of the DYPSA algorithm, we define with reference to Fig. 2 [7]: Identification rate-The percentage of larynx cycles for which exactly one GCI is detected; Miss rate-The percentage of larynx cycles for which no GCI is detected; False alarm rate-The percentage of larynx cycles for which more than one GCI is detected; Identification error, ζ - The timing error between the reference GCIs and the detected GCIs in the cycles for which exactly one GCI has been detected; Identification accuracy, σ - the standard deviation of ζ .

These metrics give us a measure of the performance of DYPSA for the instances of glottal closures in only voiced speech. We define a metric for the non-voiced regions of speech by considering the number of GCIs that are detected incorrectly in unvoiced or silence regions per second of unvoiced speech and silence. The improvement of the modified algorithm over the original DYPSA for the spurious GCIs in non-voiced speech is defined as $Q = \frac{v_{orig} - v_{\mathrm{mod}}}{v_{orig}} \times 100\%$ where v_{orig} and v_{mod} are the number of spurious GCIs detected in unvoiced and silence periods of the signal by the original DYPSA algorithm and the modified algorithm respectively.

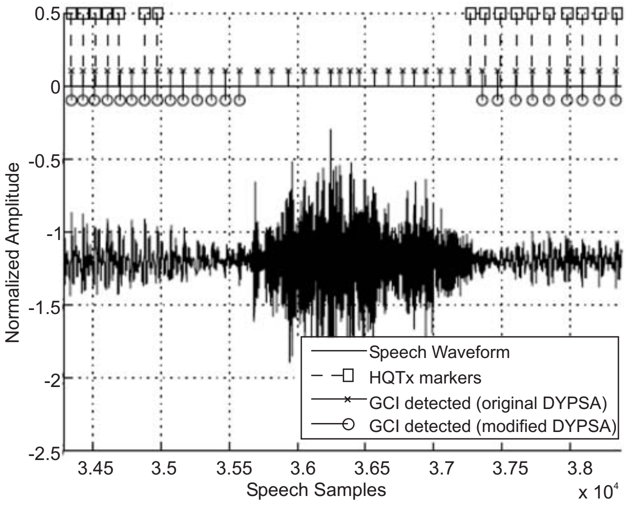

Fig. 3. Comparison between GCI detected with modified and original DYPSA

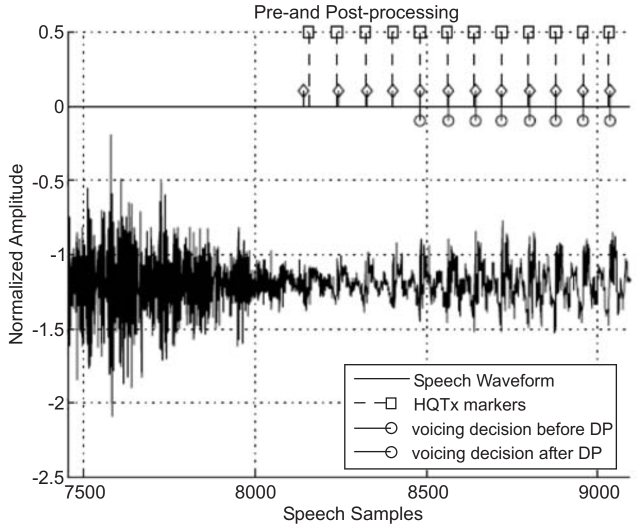

Fig. 4. GCI detection comparing pre and post- processing

Fig. 3 depicts an example of the modified DYPSA's operation. For this utterance extract, the dashed lines marked with □ indicate the true GCIs from the HQTx, the solid lines marked with × indicate the GCIs from the original version of the algorithm and the lower lines marked with ○ indicate the GCIs from our modified DYPSA algorithm. It is observed that DYPSA's GCIs match well with the EGG-derived GCIs at the start and end of the extract. The original algorithm generates spurious GCIs during the non-voiced speech. It can also be seen that our modified algorithm generates more candidates than the HQTx at the boundary from voiced to unvoiced speech transition between 3.50 and 3.55s. This is attributed to the uncertainty in voiced/unvoiced classification at the transition boundaries and, in any case, can be controlled by adjustment of the classification threshold in our method. For this example the improvement of modified DYPSA over original DYPSA is 87.7%.

It is also observed that introducing the voicing decision prior to the DP step reduces the identification rates as DYPSA misses GCIs near the onset of voiced regions due to the use of consistency measures in the cost function. From the cost functions presented in [7], the pitch deviation cost and the speech waveform similarity cost have been defined as functions of the current and previous GCI candidates under consideration by the DP. Pre-processing rejects the GCI candidates that occur in the non-voiced regions; hence we obtain misses at the beginning of voiced segments. In order to improve the detection rates, implementing the voicing decision as a post-processing (instead of pre-processing) step was investigated. Once the DP has identified a set of GCIs (for both voiced and non-voiced speech), we compute the logarithmic voicing score for each of the GCIs. The GCIs identified as occurring in the voiced speech are selected as being the true GCIs. Fig. 4 illustrates an onset of voiced speech. GCIs from HQTx are shown by the dashed lines marked with □. The solid lines marked with ○ show the results from our modified algorithm when the voicing decision is applied before the DP. The solid lines marked ◊ show the results when the voicing decision is applied as a post-processor, for which improved detection can be observed.

Table 1 shows comparative results on the APLAWD database for identification rate, miss rate, false alarm rate and the improvement of modified DYPSA over the original DYPSA with the voicing decision implemented as pre- and post-processing. We observe an improvement of 87.23% in reduction of spurious GCIs by pre-processing. Post-processing gives us reasonably close performance and the improvement in reduction of GCIs is 85.23% over the original DYPSA. We also note an increase in miss rate which is attributed to occasional misses within the voiced speech due to mixed voiced/unvoiced phonemes and misses at voicing onset/endpoint boundaries. However the transition areas are normally less problematic as the speech data at the voiced non-voiced boundaries is less useful for speech analysis.

Table 1. Performance comparison for GCI detection with voicing discrimination

	Voiced			**Non-voiced (Unvoiced & Silence)**
	Ident. Rate (%)	Miss Rate (%)	False Rate (%)	Improv-ement *Q* (%)
DYPSA[original]	95.6	1.8	2.6	0
DYPSA[modified with Pre proc.]	93.8	4.2	2.0	87.2
DYPSA[modified with Post proc.]	94.3	3.5	2.2	85.2

Conclusion

We have presented a modification of the DYPSA algorithm to include voicing discrimination that reduces the number of spurious GCIs detected in unvoiced speech or silence. The improvement is conditioned by the need to maintain the performance of DYPSA for the voiced speech segments. The technique adopted classifies a speech segment as voiced, unvoiced or silence on the basis of feature measurements extracted from the speech signal alone. For each of the candidates we obtain a normalized voicing score and identify the voiced GCI candidates. Having identified a subset of voiced GCI candidates the technique of Dynamic Programming is used for the selection of true GCIs. Having identified a subset of voiced GCI candidates, DP is used for the selection of true GCIs. Incorporating the voicing discrimination improves the detection of spurious GCIs in unvoiced segments by approximately 87% while the identification rate for voiced segments is only reduced by 1 to 2%, with most of the errors occurring in the regions of voicing onset and endpoints. Application of the voicing discrimination as both a pre- and post-processor to the DP has been studied. The post-processing approach shows slightly better identification rate for voiced speech but with slightly less improvement in the rejection of spurious GCIs in non-voiced speech. The enhanced robustness of the modified algorithm, which reduces the number of spurious GCIs, enables the use of DYPSA autonomously over entire speech utterances without the need for separate labelling of voiced regions. The ability of DYPSA to correctly identify the glottal closure instances enables the use of speech processing techniques such as close-phase LPC analysis and closed-phase glottal inverse filtering with many diverse applications in speech processing.

REFERENCES

1. A. Neocleous and P. A. Naylor, “Voice source parameters for speaker verification”, *Proceedings of the European Signal Processing Conference*, 1998, pp. 697-700.
2. D. M. Brookes and D. S. Chan, “Speaker characteristics from a glottal airflow model using glottal inverse filtering,” *Proceedings of Institute of Acoustics*, Vol. 15, 1994, pp. 501-508.
3. B. Atal, “Predictive coding of speech at low bit rates”, *IEEE Transactions on Communications*, Vol. 30, No. 4, Apr. 1982, pp. 600-614.
4. A. S. Spanias, “Speech coding: A tutorial review,” *Proceedings of the IEEE*, Vol. 82, No. 10, Oct. 1994, pp. 1541-1582.
5. J. H. Eggen, “A glottal excited speech synthesizer,” *IPO Annual Progress Report*, 1989.
6. F. Charpentier and E. Moulines, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Proceedings of EUROSPEECH*, Vol. 2, 1989, pp. 13-19.
7. P. A. Naylor, A. Kounoudes, J. Gudnason and M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.
8. R. Smits and B. Yegnanarayana, “Determination of instants of significant excitation in speech using group delay function,” *IEEE Transactions on Speech Audio Processing*, Vol. 3, 1995, pp. 325-333.
9. R. Schwartz and Y.-L. Chow, “The N-best algorithm: An efficient and exact procedure for finding the n most likely sentence hypotheses,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 81-84.
10. J. K. Chen and F. K. Soong, “An N-best candidates-based discriminative training for speech recognition applications”, *IEEE Transactions on Speech Audio Processing*, Vol. 2, Jan 1994, pp. 206-216.
11. G. Lindsey, A. Breen, and S. Nevard. “SPAR’S archivable actual-word database”, *Dept. Phonetics and Linguistics, University College London, Technical Report*, June 1987.
12. D. Y. Wong, J. D. Markel, and J. A. H. Gray, “Least squares glottal inverse filtering from the acoustic speech waveform”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 27, Aug 1979, pp. 350-355.
13. C. Ma, Y. Kamp, and L. F. Willems, “A Frobenius norm approach to glottal closure detection from the speech signal,” *IEEE Transactions on Speech Audio Processing*, Vol. 2, Apr. 1994, pp. 258-265.
14. B. Atal and L. Rabiner, “A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 24, No. 3, Jun. 1976, pp. 201- 212.
15. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, New Jersy: Prentice Hall, 1993.
16. L. Siegel and K. Steiglitz, “A pattern classification algorithm for the voiced/unvoiced decision,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, April 1976, pp. 326-329.
17. A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, Series B, Vol. 39, No. 1, 1977, pp. 1-38.
18. Teknomo, Kardi. “K-Means Clustering Tutorials,” http:\\people.revoledu.com\kardi\ tutorial\kMean\
19. G. Singh, A. Panda, S. Bhattacharyya and T. Srikanthan, “Vector quantization techniques for GMM based speaker verification”, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, April 2003.
20. T. K. Moon, “The expectation-maximization algorithm,” IEEE Signal Processing Magazine, Vol. 13, No. 6, Nov. 1996, pp. 47-60.
21. M. Huckvale, “Speech Filing System: Tools for Speech Research”, University College London, 2000, http://www.phon.ucl.ac.uk/resource/sfs/