ENHANCED ROBUSTNESS TO UNVOICED SPEECH AND NOISE IN THE DYPSA ALGORITHM FOR IDENTIFICATION OF GLOTTAL CLOSURE INSTANTS

Hania Maqsood[1], *Jon Gudnason*[2], *Patrick A. Naylor*[2]

[1]Bahria Institue of Management and Computer Sciences, Islamabad, Pakistan
[2]Imperial College, Department of Electrical and Electronic Engineering, Exhibition Road, London, UK
email: p.naylor@imperial.ac.uk

ABSTRACT

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

1. INTRODUCTION

The classical model of the human speech production system generally comprises a linear vocal 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. Separation of the vocal tract from the source effects is usually based on accurate estimations of glottal closure instants (GCIs) and the use of larynx synchronous processing techniques such as closed-phase 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 new modified version of the DYPSA algorithm which maintains all the advantages of DYPSA's high accuracy in voiced speech but overcomes a problem with the original form of the algorithm during unvoiced speech in which spurious GCIs are erroneously detected. This is to be achieved by estimating the likelihood that each GCI occurs within voiced speech and suppress any GCIs for which this likelihood is below a determined threshold.

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.

2. 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 the phase-slope function as presented in [8]. The GCIs are identified from this phase-slope function as positive-going zero-crossings. DYPSA also involves identification of additional candidates using the technique of phase-slope projection [7]. An optimum set of GCIs is then selected from the candidates by minimizing a cost function using N-best Dynamic Programming (DP) [9][10]. The cost function comprises five components: 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 APLAWD speech database [11] with the reference GCIs extracted from the EGG signal. A comparative evaluation of DYPSA with 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, detects spurious GCIs for unvoiced 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 voiced/unvoiced classification problem. The solution to this classification problem involves incorporating a voicing decision for the GCI candidates within the algorithm. The GCI candidates identified as occurring in the unvoiced speech are then removed.

2.1 Identification of GCI Candidates

The speech signal with sampling frequency 20 kHz is passed through a 1st order pre-emphasis filter with a 50 Hz cut-off frequency and processed using autocorrelation LPC 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 the vicinity of each sample n as:

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

with Fourier transform $X_n(\omega)$. The phase slope func-

tion [8]

$$\tau_n(\omega) = \frac{d\arg(X_n(\omega))}{d\omega} \tag{2}$$

is defined as the average slope of the unwrapped phase spectrum of the short time Fourier transform of the linear prediction residual. DYPSA identifies GCIs as 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 occasionally 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. 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 illustrated 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.

2.2 Dynamic Programming

The selection of true GCIs from the set of GCI candidates is performed by minimizing a cost function using N-best DP [9][10]. The N-best DP procedure maintains information about the N most likely hypotheses at each step of the algorithm. The value of N has been chosen as 3 following the approach in [7]. The cost function to be minimized by DP is

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \lambda^T \mathbf{c}_{\Omega}(r) \tag{3}$$

where the weights are obtained using an optimization procedure [7] as

$$\lambda = [\lambda_A\ \lambda_P\ \lambda_J\ \lambda_F\ \lambda_S]^T = [0.8\ 0.5\ 0.4\ 0.3\ 0.1] \tag{4}$$

and Ω is a subset of GCIs selected from all GCI candidates, $|\Omega|$ is the size of Ω, r indexes the elements of Ω and ^{T} represents the transpose operator.

The elements of the cost vector evaluated for the r^{th} GCI of subset Ω are

$$\mathbf{c}_{\Omega}(r) = [c_A(r), c_P(r), c_J(r), c_F(r), c_S(r)]^T \tag{5}$$

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 DP then searches for the subset of GCIs giving minimum cost. The advantage of using the DP cost function is that it effectively penalizes GCI candidates in a way such that in most cases all but one candidate per larynx cycles is rejected. The reader is referred to [7] for further details.

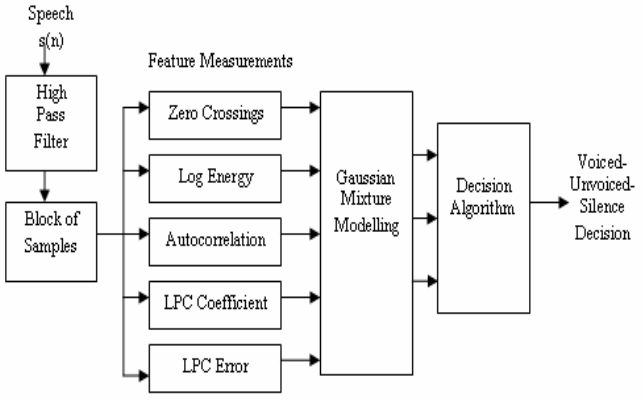

Figure 1: Block diagram of voiced-unvoiced-silence detector.

3. 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 the signal where no speech is present and generally contains at least some level of background noise. The technique adopted for speech classification takes into consideration the statistical distributions and characteristic 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.

3.1 Feature Extraction

The speech signal is initially high-pass filtered at approximately 200 Hz. Frames of duration 10 ms are then defined centred on each GCI candidate found from DYPSA using the procedure described in Section 2.1 and features are then extracted for each frame. The choice of the features 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 relatively low zero-crossing rate while unvoiced speech has a concentration of energy at high frequencies and therefore typically exhibits a higher zero-crossing rate. The zero-crossing rate for silence depends 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 is a small positive constant added to prevent computing log of zero. In moderate or good noise conditions, the energy of voiced sounds is significantly higher than the energy of unvoiced speech or silence.

3) Normalized Autocorrelation Coefficient is defined as

$$C_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{\sum_{n=1}^{N} s^2(n)\sum_{n=0}^{N-1} s^2(n)}}. \tag{7}$$

Adjacent samples of voiced speech are highly correlated, therefore C_1 is close to unity, whereas for unvoiced speech, the correlation is closer to zero.

4) First Predictor Coefficient from Linear Predictive Analysis. It was shown by Atal [14] that the first predictor coefficient is identical (with a negative sign) to the cepstrum of the log spectrum at unit sample delay. Therefore the first LPC coefficient can be used to help to discriminate between the three classes of signal, each of which has differing spectral characteristics evident in the first predictor coefficient.

5) Normalized Prediction Error. The normalized prediction error from linear prediction can be written in dB [15] as

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

where E_s is given in (6) and $\phi(i,k) = \frac{1}{N}\sum_{n-1}^{N} s(n-i)s(n-k)$ is the (i,k) element of the covariance matrix of the speech signal. The normalized prediction error is large at glottal closures in voiced speech since the voiced excitation cannot be well represented by the AR model employed in the predictor.

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 combine their contributions in discriminating between the three classes.

3.2 Gaussian Mixture Modelling

It is assumed that the above features are from a multidimensional Gaussian distribution where each class is modelled as a Gaussian-shaped cluster of points in five-dimensional feature space. This assumption has the advantages of computational simplicity as the decision rule is determined by the mean vector μ and covariance matrix C. In order to estimate the parameter set we employ the K-means clustering algorithm followed by iterations of the Expectation Maximization (EM) algorithm. The K-means algorithm [17][18] partitions the points of a data matrix into K clusters. The EM algorithm [19][20] then maximizes the log-likelihood from data in order to estimate the parameters of the distribution. 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.

3.3 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 and covariance matrix C_i is given by

$$g_i(x) = (2\pi)^{-d/2}|C_i|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_i)^T C_i^{-1}(x-\mu_i)\right) \quad (9)$$

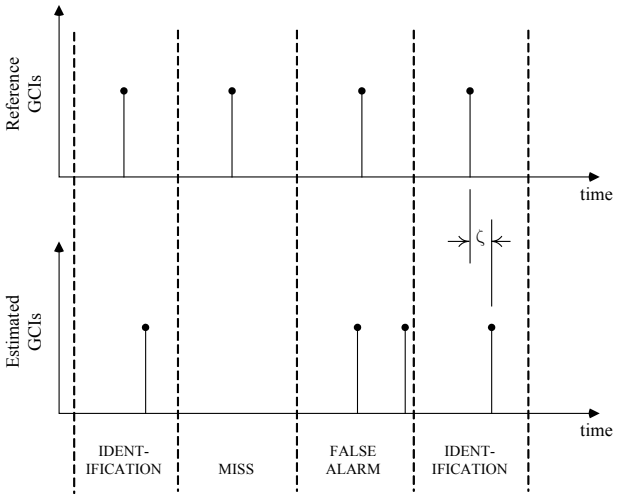

Figure 2: Definition of evaluation metrics. The dotted lines depict a frame defined around each reference GCI marker to indicate a larynx cycle (after [7]).

where $|C_i|$ 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, we work in the log domain. Taking the natural log on both sides of (9) we obtain

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

from which we can 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 non-voiced 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 empirically as suitable for the APLAWD database. GCI candidates with scores below this threshold are excluded from further processing. This avoids DYPSA giving spurious GCIs during unvoiced speech or silence and also simplifies the computation required for the DP routine within DYPSA.

4. EXPERIMENTS AND RESULTS

For the performance comparison of the original DYPSA algorithm and our proposed modified version, we require reference GCIs which are obtained from time-aligned simultaneously recorded EGG signals in the APLAWD database. Reference GCIs are then extracted from the EGG signal using HQTx algorithm [21]. The HQTx markers (indicating

'ground truth' GCIs in the speech waveform) are then compared to the GCIs obtained from DYPSA using (i) *Identification rate* - the percentage of larynx cycles for which exactly one GCI is detected; (ii) *Miss rate* - the percentage of larynx cycles for which no GCI is detected; (iii) *False alarm rate* - the percentage of larynx cycles for which more than one GCI is detected; (iv) *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; and (v) *Identification accuracy* - the standard deviation of ζ. These terms are illustrated in Fig. 2 [7].

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 $Q = \frac{v_{orig} - v_{mod}}{v_{orig}} \times 100\%$ where v_{orig} and v_{mod} are the number of spurious GCIs detect in unvoiced and silence periods of the signal by the original DYPSA algorithm and the modified algorithm respectively.

Fig. 3 depicts an example of the modified DYPSA's operation. For this utterance extract, the dashed lines marked with $\square$ indicate the true GCIs from HQTx, the solid lines marked with $\times$ indicate the GCIs from the original version of the DYPSA algorithm and the lower solid lines marked with $\circ$ indicate the GCIs from our modified DYPSA algorithm. It is observed that DYPSA's GCIs match well in general with the EGG-derived GCIs from HQTx during the voiced regions at the start and end of this extract. The original DYPSA algorithm generates spurious GCIs during the unvoiced region at the centre of the extract whereas our modified DYPSA algorithm doesn't generate spurious GCIs during the unvoiced regions. It can also be seen that our modified algorithm generates more candidates than HQTx at the boundary from voiced to unvoiced speech between 3.50 and 3.55 s. This is explained by the uncertainty in voiced/unvoiced classification at voicing 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 when running tests over the complete APLAWD database that introducing the voicing decision prior to the DP step reduces the identification rate as DYPSA misses GCIs near the onsets and endpoints 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 function and the speech waveform similarity cost are defined as a function of the current and previous GCI candidates under consideration by the DP. Pre-processing rejects the GCI candidates that occur in the unvoiced regions, hence causing misses at the boundaries of some voiced segments. In order to improve the detection rate, implementation of 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 $\square$. The solid lines marked $\circ$ show

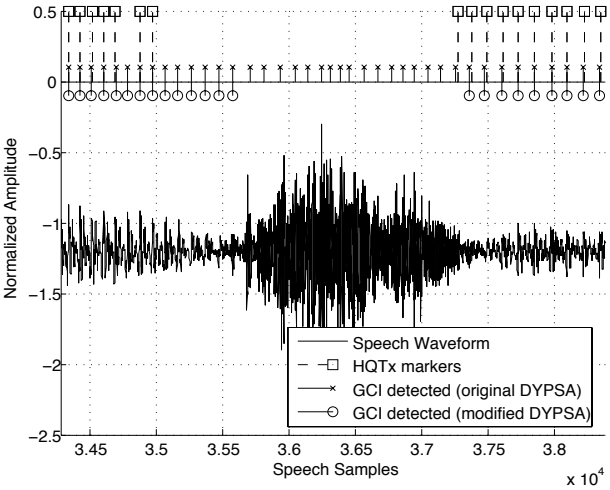

Figure 3: GCI detection with modified DYPSA.

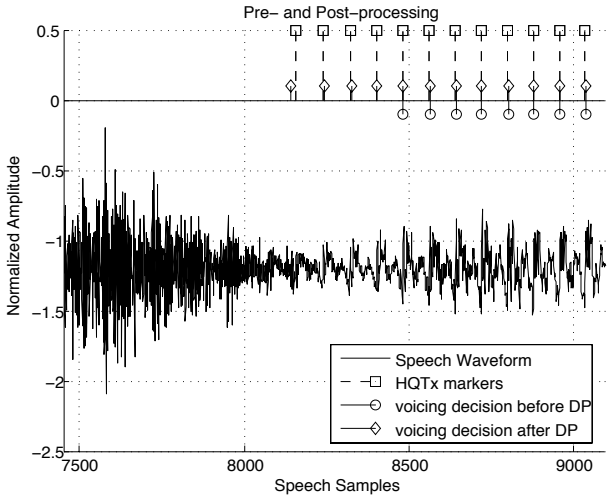

Figure 4: GCI detection with modified DYPSA comparing pre- and post-processing.

the results from our modified algorithm when the voicing decision is applied as a pre-processor to the DP. The solid lines marked $\diamond$ show the results when the voicing decision is applied as a post-processor, for which improved detection can be observed.

Table I shows comparative results on the APLAWD database for identification rate, miss rate, false alarm rate and the improvement over the original DYPSA with the voicing decision implemented as pre- and post-processing. We observe an improvement of 87.2% in the detection of spurious GCIs using pre-processing compared to original DYPSA on the APLAWD database. Post-processing achieves an 85.2% improvement. 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, such misses are usually of low importance since speech data near onsets and endpoints is often less useful for speech analysis.

Table 1: Performance comparison for GCI detection with voicing discrimination.

	Voiced			Unvoiced
	Ident. Rate (%)	Miss Rate (%)	False Rate (%)	Improvement Q (%)
DYPSA	95.6	1.8	2.6	0
DYPSA Pre-proc.	93.8	4.2	2.0	87.2
DYPSA Post-proc.	94.3	3.5	2.2	85.2

5. 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 obtained is conditioned by the need to maintain the high performance of DYPSA for voiced speech. 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, 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 unvoiced 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," in *Proc. 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," *Proc. Institute of Acoustics*, vol. 15, pp. 501–508, 1994.

[3] B. Atal, "Predictive coding of speech at low bit rates," *IEEE Transactions on Communications*, vol. 30, no. 4, pp. 600–614, Apr 1982.

[4] A. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[5] J. H. Eggen, "A glottal excited speech synhesizer," *IPO Annual Progress Report*, 1989.

[6] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Proc. 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 Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[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, Sep 1995.

[9] R. Schwartz and Y.-L. Chow, "The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. IEEE Int. Conf. Acoust., Speech, 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 Trans. Speech Audio Processing*, vol. 2, pp. 206–216, Jan 1994.

[11] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Tech. Rep., Jun 1987.

[12] D. Y. Wong, J. D. Markel, and J. A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 350–355, Aug 1979.

[13] C. 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, Apr 1994.

[14] B. Atal and L. Rabiner, "A pattern recogniion approach to voice-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. ASSP*, vol. 24, no. 3, pp. 201–212, Jun. 1976.

[15] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[16] L. Siegel and K. Steiglitz, "A pattern classification algorithm for the voiced/unvoiced decision," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Apr 1976, pp. 326–329.

[17] K. Teknomo, "K-means clustering tutorials," [Online] http://people.revoledu.com/kardi/tutorial/kMean.

[18] G. Singh, A. Panda, S. Bhattacharyya, and T. Srikanthan, "Vector quantization techniques for gmm based speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 6-10 April 2003, pp. II–65–8vol.2.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Roal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[20] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, Nov. 1996.

[21] M. Huckvale, *Speech Filing System: Tools for Speech Research*, University College London, 2000, [Online] http://www.phon.ucl.ac.uk/resource/sfs/.