EVALUATION OF PITCH ESTIMATION IN NOISY SPEECH FOR APPLICATION IN NON-INTRUSIVE SPEECH QUALITY ASSESSMENT

Dushyant Sharma, Patrick. A. Naylor

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

ABSTRACT

Pitch estimation has a central role in many speech processing applications. In voiced speech, pitch can be objectively defined as the rate of vibration of the vocal folds. However, pitch is an inherently subjective quantity and cannot be directly measured from the speech signal. It is a nonlinear function of the signal's spectral and temporal energy distribution. A number of methods for pitch estimation have been developed but none can claim to work accurately in the presence of high levels of additive noise or reverberation. Any system of practical importance must be robust to additive noise and reverberation as these are encountered frequently in the field of operation of voice telecommunications systems. In non-intrusive speech quality measurement algorithms, such as the P.563 and LCQA, pitch is used as a feature for quality assessment. The accuracy of this feature in noisy speech signals will be shown to correlate with the accuracy of the objective measure of the quality of the speech signal. In this paper we evaluate the performance of four established state-of-the-art algorithms for pitch estimation in additive noise and reverberation. Furthermore, we show how accurate estimation of the pitch of a speech signal can influence objective speech quality measurement algorithms.

1. INTRODUCTION

Pitch estimation has an important role in a number of applications, including speech synthesis, recognition and as metadata in multimedia applications [1]. It is also used as a feature in many objective speech quality assessment algorithms such as the P.563 and the LCQA algorithms. The area of pitch estimation has attracted a lot of interest resulting in a number of algorithms for pitch estimation. However, none of the current algorithms has the desired robustness to noise and reverberation, degrading their usefulness in many potential algorithms, such as objective speech quality assessment.

Pitch detection in speech signals may be described as the accurate estimation of the perceived tone of a speech signal. The perceived pitch of a speech signal is an inherently subjective quantity which correlates well with the fundamental frequency of the signal [2]. Pitch tracking algorithms aim to estimate the inverse of the smallest true period in the interval of interest. However, estimation of the fundamental frequency of a speech signal from the speech waveform alone is a challenging problem due to the quasi-periodic nature of pitched speech and mixed nature of the excitation [3].

Pitch arises due to the oscillation of the vocal folds which modulates the airflow through the glottis. This modulation of the airflow serves as the excitation for the vocal tract during voiced speech. Pitch plays an important role in contributing to the prosody in human speech as well as distinguishing segmental categories in tonal languages.

One of the objectives of this paper is to highlight the importance of pitch estimation robustness in nonintrusive speech quality assessment algorithms such as the Low-Complexity, Nonintrusive Speech Quality Assessment algorithm (LCQA) [4].

2. PITCH TRACKING ALGORITHMS

This section describes the four algorithms used for the comparative evaluation of pitch tracking in noise and reverberation. Also, described is the SIGMA algorithm, which was used to obtain a ground-truth reference in the form of glottal closure instants (GCIs) from the laryngograph recording (EGG).

2.1 Robust Algorithm for Pitch Tracking (RAPT)

RAPT [2] is a frame based algorithm which uses normalized cross correlation (NCCF) (1) as the primary candidate generation function and uses dynamic programming to refine the pitch estimation. The NCCF, $\phi_{i,k}$ (for lag k and analysis frame i) is the autocorrelation function normalized by the energy of the input signal defined as

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}}, k=0,...,K-1; m=iw; i=0,...,M-1, \quad (1)$$

where,

$$e_j = \sum_{l=j}^{j+n-1} s_l^2, \quad (2)$$

where the number of samples in each window is n and the frame is advanced at each iteration by w samples. The input signal s is assumed to be zero mean.

The NCCF is the most computationally expensive operation in RAPT and so the algorithm performs the NCCF in a two pass process. A down-sampled version of the input signal is used to estimate the first set of candidate peaks, followed by a high resolution (full sample rate) NCCF around the candidates of interest. The algorithm is summarized below:

- Periodically compute the NCCF of the down sampled signal for all lags in the range of pitch. Locations of local maxima in this 1st pass of the NCCF are recorded.
- Compute the high resolution NCCF (signal at original sampling frequency) only around the peak locations recorded in previous step.
- Search for local maxima in the high resolution NCCF to obtain improved peak locations and amplitude estimates.
- Dynamic Programming [5] is used to select the set of NCCF peaks or unvoiced hypothesis across all frames.

The Voicebox [6] implementation of this algorithm was used for the comparative evaluation of RAPT.

2.2 P.563 Pitch Detection Module

This is the pitch estimator used in the ITU-T P.563 [7] objective speech assessment algorithm and is also based on the autocorrelation function. The autocorrelation is calculated over 65 ms frames with 50 percent overlap in the frequency domain as

$$R_{xx}(t) = \int_{-\infty}^{\infty} Y(\omega)Y^*(\omega)e^{j\omega t}d\omega, \qquad (3)$$

where $Y(\omega)$ is the Discrete Fourier Transform (DFT) of the signal. In practice a Fast Fourier Transform (FFT) is applied. The autocorrelation R_{xx} is normalized by $R_{xx}(0)$. The algorithm then searches within a range of lags of interest for a maximum after filtering the signal through a Hanning window and performs some post-processing to avoid pitch doubling.

2.3 YIN Pitch Tracker

The YIN [8] algorithm uses a difference function based on the autocorrelation function as the candidate generator in conjunction with a number of optimization steps. Named after the oriental *yin-yang* principle of duality, it aims to balance between the autocorrelation and the cancelation that it involves. The algorithm's main processing blocks are described below:

- Difference Function (DF) (5) - this is the candidate generation function used in YIN. While the autocorrelation function aims to maximize the product between the waveform and its delayed duplicate, the difference function aims to minimize the difference between the waveform and its delayed duplicate. The underlying assumption is that the difference between a periodic signal x_t of period T and its time shifted version x_{t+T} is 0, i.e.

$$\sum_{j=t+1}^{t+W} (x_j - x_{j+T})^2 = 0. \qquad (4)$$

  This assumption holds true after taking the square and averaging over a window (4). The unknown period may be found by searching in the window for the value of τ which makes the difference function,

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2 \qquad (5)$$

  equal to zero.
- Cumulative mean normalized difference function - in order to handle the quasi-periodic nature of pitch, the YIN algorithm normalizes the DF by its cumulative mean and sets a value of 1 for $\tau = 0$, as

$$d_t'(\tau) = \begin{cases} 1, & \text{if } \tau = 0 \\ d_t(\tau)/[(1/\tau)\sum_{j=1}^{\tau} d_t(j)] & \text{otherwise.} \end{cases} \qquad (6)$$

- Absolute Threshold, Parabolic Interpolation and Local Search - the last three steps involve placing a threshold on the smallest value of τ that is accepted. Also, parabolic interpolation is used to refine the peak location and searching around initial pitch markers to further refine the estimate.

2.4 Dynamic Programming Projected Phase-Slope Algorithm (DYPSA)

The DYPSA [9] algorithm was originally designed for automatic estimation of glottal closure instants (GCIs) in voiced speech but as a consequence also gives pitch information. The algorithm is based on an enhancement of the group delay algorithm [10] by R. Smiths and B. Yegnanarayana, which is used as the primary candidate generator. DYPSA uses dynamic programming (DP) to identify the best GCI candidates by minimizing some cost functions. The DYPSA algorithm operates on the speech signal alone and does not require an EGG reference signal. The pitch estimate is derived from the inter GCI duration and mapped into frames.

2.5 SIGMA Algorithm for Glottal Activity Detection in EGG signals

The SIGMA [11] algorithm operates on an EGG signal and identifies the glottal closure instants (GCIs) and glottal opening instants (GOIs) for voiced speech. It has been used here to obtain reference GCIs from the contemporaneous EGG signal available in the database used for evaluation and provides the ground truth in the evaluation.

The SIGMA algorithm is based on a stationary wavelet transform preprocessor, with a group delay function as the peak detection function. Gaussian Mixture Modeling is used to classify true and false detections to further improve the performance of the algorithm. The SIGMA algorithm has been shown to provide an average GCI hit rate greater than 99% [11] when compared to hand-labeled GCIs.

The period between two consecutive GCI's is taken as the pitch period, which is then mapped into frames as with the DYPSA algorithm for evaluation with other pitch estimation algorithms.

3. EVALUTION

The first part of this paper concentrates on the evaluation of four established algorithms under noise and reverberation. Two classes of acoustic degradation were considered:

- Additive Noise - this is most commonly perceived as 'background' noise. For this evaluation, car, babble and white noise were used. Signal-to-noise ratios of -10, 0, 10, and 20 dB were used to represent the entire range of the speech signal degradation.
- Reverberation - the method of images [12, 13] was used to generate the impulse response of a rectangular room (length 5m, width 4m, height 3m) with reverberation times (T_{60}) of 0.1, 0.3 and 0.5 seconds. In addition to the isolated additive noise and reverberation tests, a set of tests were carried out by combining the effects of T_{60} = 0.1s reverberation to 10 dB SNR speech signals to represent multiple degradations.

The SAM database [14] of English speech was used for the evaluation. It contains 2 male and 2 female speakers and also has contemporaneous recordings of laryngograph signals for the spoken sentences. The SIGMA [11] algorithm was used for the extraction of glottal closure instants (GCIs) and map them to an estimate of the pitch period by considering the time between two GCIs as the pitch period. Then the pitch period was interpolated into frames of size dictated by the pitch estimation algorithm being tested and converted to

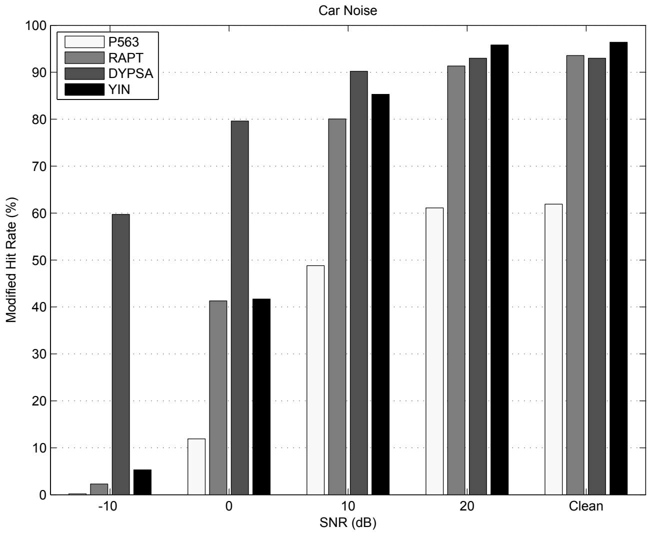

Figure 1: Pitch estimation in car noise with SNR (x-axis) from -10 dB (left) to clean speech (right). Performance metric is the modified hit rate (y-axis) given as a percentage of overall hits.

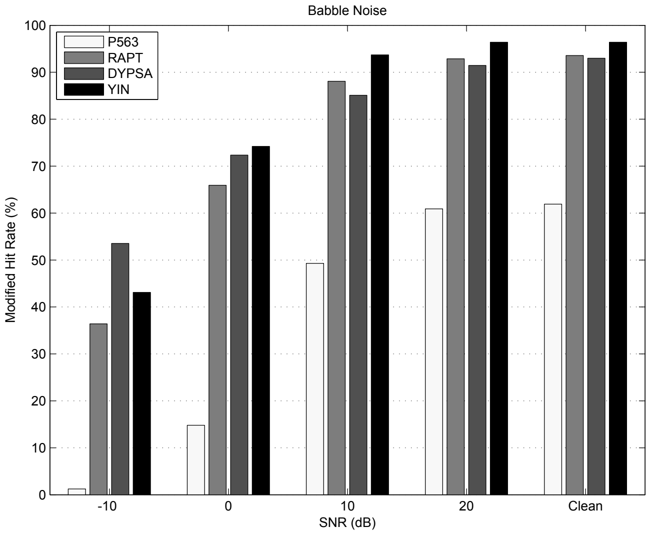

Figure 2: Pitch estimation in babble noise with SNR (x-axis) from -10 dB (left) to clean speech (right). Performance metric is the modified hit rate (y-axis) given as a percentage of overall hits.

pitch per frame. This formed the ground truth for the evaluation of the pitch estimators.

We define two measures for the purpose of this evaluation as follows. Accuracy is defined as the root mean square (RMS) difference between the true pitch period in a frame i (T_i) and the estimated pitch period ($\hat{T}_i$). A hit is defined as a pitch mark occurring in a frame for which the ground truth, obtained through SIGMA, also placed a pitch mark in the frame of interest. The analysis is restricted to voiced regions of the signal as obtained from the SIGMA algorithm.

Our overall measure is then defined as the modified hit rate (MHR), which is a hit with an accuracy of 80% and higher as

$$MHR = \frac{\sum(hits\ with\ accuraccy >= 80\%)}{no.\ of\ voiced\ frames} \times 100. \qquad (7)$$

The advantages of this evaluation methodology are that a meaningful interpretation can be made of the performance of different pitch tracking methods in terms of the number of 'good' hits - where 'good' here is defined for accuracy greater than 80%. We note that our methodology is a straightforward development of the combination of methodologies employed in [9] and [11].

4. EXPERIMENTS AND RESULTS

4.1 Pitch Tracking Experiments

We present the results obtained from the evaluation of the four pitch estimation algorithms in noise, reverberation and noise and reverberation.

Figure 1 shows how the four algorithms perform in car noise. We can see that at 20 dB SNR, the performance in terms of the modified hit rate (MHR) is close to the performance achieved in clean speech for all algorithms. However, for the lower SNR's of 0 and -10 dB, all dedicated pitch tracking algorithms fail, as shown by the low MHR score, even DYPSA provides a significantly lower MHR. A similar result is obtained for the case of babble noise as shown in Fig. 2. In the presence of white noise, both DYPSA and RAPT perform poorly at low SNR's. However, the YIN algorithm performs well even at an SNR of 0 dB, achieving an MHR of 93.1%, as shown in 3. The P.563 pitch tracking module performs poorly in all noise conditions, this can be explained by the low complexity and simplicity of the algorithm, suggesting that the P.563 algorithm is not very sensitive to the correctness of its pitch tracking module.

In the case of reverberation, we can see from Fig. 4 that both RAPT and YIN perform well in reverberation, achieving an MHR of 68.3% and 79.8% respectively in a highly reverberant room (T_{60} = 0.5 s). However, the DYPSA algorithm is seen to be more sensitive to reverberation.

From Fig. 5 we can see the effect of 10 dB SNR of additive noise in a reverberant room with T_{60} = 0.1 s. It is clear that all the four algorithms fail to work in a slightly reverberant room with a small amount of additive noise. However, when only one degradation is present, RAPT, YIN and DYPSA perform well in those conditions.

4.2 Speech Quality Assessment Experiments

We next consider what effect pitch estimation errors have on speech quality assessment. In the context of non-intrusive speech quality assessment, important measures include the ITU-T P.563 [7] measure and the LCQA algorithm [4]. This paper will focus on the LCQA approach.

- Frame the input speech signal for further processing
- Derive the per frame features, including the pitch period and its first time derivative
- Build a statistical description from the per frame features using their mean, variance and skewness properties, yielding a global feature set

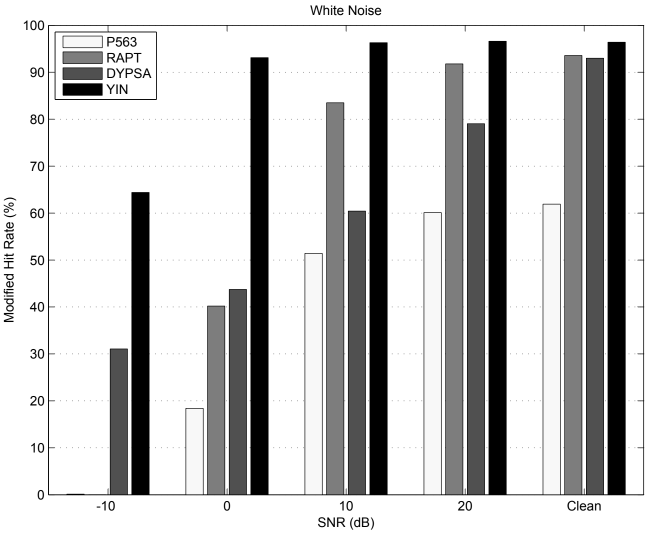

Figure 3: Pitch estimation in white noise with SNR (x-axis) from -10 dB (left) to clean speech (right). Performance metric is the modified hit rate (y-axis) given as a percentage of overall hits.

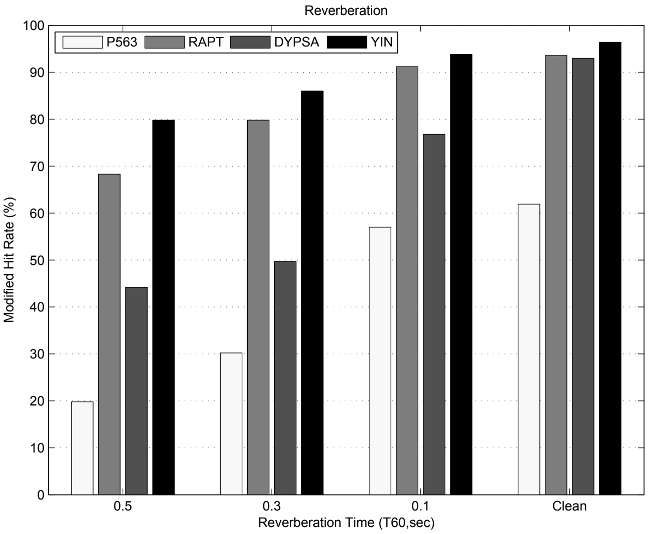

Figure 4: Pitch estimation in a reverberant room of length 5m, width 4m, height 3m. Reverberation time (x-axis) T_{60}= 0.5 s (left) to clean speech in a non-reverberant room (right). Performance metric is the modified hit rate (y-axis) given as a percentage of overall hits.

Table 1: Correlation Coefficients for testing and and training of LCQA on entire P.23 database with RAPT and YIN pitch estimation algorithms.

	RAPT	YIN
Correlation coefficient (R)	0.6091	0.6532

Gaussian Mixture Modeling (GMM) is then used to infer the speech quality of the input signal based on this feature set and a previously trained GMM. The LCQA algorithm is data driven and requires a GMM to be trained. The performance of the GMM-based probability mapping depends on the amount of training data available. For our evaluation, the English subset of 176 speech files from the P.23 [15] database were used, out of which 136 were used for training with 6 mixtures. The testing was done on the remaining 40 speech files from the English subset. The P.23 database contains subjective mean opinion scores (MOS) for a range of degraded speech samples. The RAPT and YIN pitch trackers were used in both training and testing phases of the LCQA and the metric used for comparison of performance was the correlation coefficient R, defines as

$$R = \frac{\sum_i(\hat{Q}_i - \mu_{\hat{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_i(\hat{Q}_i - \mu_{\hat{Q}})^2 \sum_i(Q_i - \mu_Q)^2}}, \quad (8)$$

where $\hat{Q}$ is the estimated speech quality (also known as MOS-LQO) and Q is the subjective speech quality (also known as MOS-LQS).

Table 1 shows how using YIN, which is a more robust pitch estimation algorithm than RAPT, improves the performance of LCQA in terms of increasing the correlation coefficient between the estimated and the subjective speech quality.

5. CONCLUSIONS

The algorithms RAPT, DYPSA, YIN and P.563 Pitch Tracking Module were evaluated in terms of the modified hit rate (MHR) under various noise and reverberation conditions. It was shown that pitch tracking in additive noise alone is a challenging task, with all algorithms giving unreliable results for SNRs below 10 dB. For pitch tracking in reverberation alone, performance was poor below reverberation time of T_{60}=0.3 s. Whereas pitch tracking in modest reverberation (T_{60}=0.1 s) and additive noise (SNR 10 dB) was shown to produce extremely poor results, with average performance at 30% (MHR).

The evaluation of the different noise conditions mentioned above led to the conclusion that all four algorithms fail to achieve an 80% MHR threshold when the SNR is lower than 10 dB. For the case of reverberation, T_{60} = 0.3 s is the most reverberation that can be tolerated to achieve this threshold. Also, the YIN algorithm proved to be the most robust to noise in the 10 to 20 dB SNR range, achieving a MHR above 80%. The DYPSA algorithm has been shown to perform well in car and babble noise and may have the potential with some modifications to provide a robust estimate of pitch in noise. Also, the P.563 pitch tracker has a low performance due to its simplistic approach in estimating the pitch and thus fails in noisy conditions. The RAPT algorithm performs well in noise and reverberation separately, with a performance slightly lower than that of YIN. This is a significant result as it means that any pitch estimate obtained from a speech signal with an SNR lower than 10 dB is likely to unreliable, having serious consequences for any system that relies upon accurate estimation of the pitch of a speech signal.

Also we considered the effect of pitch tracking accuracy on non-intrusive speech quality assessment algorithms using LCQA as an example. It was shown that a correlation coef-

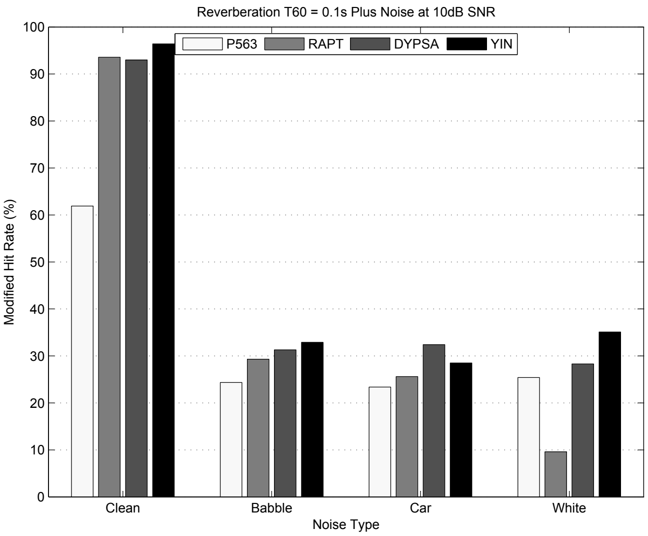

Figure 5: Pitch tracking in a reverberant room (T_{60}=0.1 s) with additive noise (10 dB SNR). Performance metric is the modified hit rate (y-axis) given as a percentage of overall hits.

ficient improvement of 0.05 was obtained by switching between RAPT and YIN in the LCQA algorithm with testing conducted on the English subset of the P.23 database [15].

Thus, the development of pitch tracking algorithms that are robust to additive noise at low SNRs and reverberation remains an important area of research with many opportunities to enhance the capabilities of current techniques.

6. ACKNOWLEDGEMENT

We would like to thank Mr. Mike Brookes for making the DYPSA algorithm available through Voicebox, and Mr. Mark Thomas for providing the Matlab implementation of the SIGMA algorithm.

REFERENCES

[1] A. de Cheveigne and H. Kawahara, "Comparative evaluation of F0 estimation algorithms," in *Proc Eurospeech*, 2001.

[2] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, pp. 495–518.

[3] L. R. Rabiner, M. J. Cheng, A. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 24, pp. 399–418, 1976.

[4] V. Grancharov, D. Zhao, J. Lindblom, and W. Kleijn, "Low-Complexity, Nonintrusive Speech Quality Assessment," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1948–1956, 2006.

[5] R. Bellman, *Dynamic Programming*. Princeton, N.J.: Princeton University Press, 1957.

[6] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[7] ITU-T, "Single-ended method for objective speech quality assessment in narrow-band telphony applications," ITU-T Recommendation P.563, 2004.

[8] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[9] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech using the DYPSA Algorithm," *IEEE Trans. Speech Audio Processing*, vol. 15, no. 1, pp. 34–43, January 2007.

[10] R. Smits and B. Yegnanarayana, "Determination of Instants of Significant Exitation in Speech using Group Delay Function," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, pp. 325–333, September 1995.

[11] M. R. P. Thomas and P. A. Naylor, "The SIGMA Algorithm for Estimation of Reference-Quality Glottal Closure Instants from Electroglottograph Signals," in *Proc. European Signal Processing Conference*, Lausanne, Switzerland, August 2008.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr 1979.

[13] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room." *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1527–1529, Nov 1986.

[14] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouronopoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeilieger, "EUROM - A Spoken Language Resource for the EU," in *Proc. European Signal Processing Conf*, September 1995, pp. 867–870.

[15] ITU-T, "ITU-T coded-speech database," ITU-T Supplement P.Sup23, Feb. 1998.