MULTICHANNEL DYPSA FOR ESTIMATION OF GLOTTAL CLOSURE INSTANTS IN REVERBERANT SPEECH

Mark R. P. Thomas, Nikolay D. Gaubitch and Patrick A. Naylor

Department of Electrical and Electronic Engineering
Imperial College London
SW7 2AZ, UK
E-mail: {mrt102, ndg, p.naylor}@imperial.ac.uk

ABSTRACT

Identification of glottal closure instants (GCIs) is important in speech applications which benefit from larynx-synchronous processing. In modern telecommunication applications, speech signals are often obtained inside office rooms, with one or more microphones placed at a distance from the talker. Such speech signals are affected by reverberation due to the reflections from surrounding walls and objects, which distort the observed speech signals and degrade the performance of speech processing algorithms.

This paper presents a study of the identifiability of GCIs from reverberant speech using the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) and new extensions to the multimicrophone case. Two multichannel algorithms are proposed and evaluated; in both cases, considerable performance gains over a single microphone are obtained, with detection rates improved by up to 29% in highly reverberant environments.

1. INTRODUCTION

Identification of glottal closure instants (GCIs) in voiced speech is important for many speech processing applications such as larynx-synchronous processing in speech synthesis [1], prosodic speech modification [2] and speech dereverberation [3]. The GCIs can be identified accurately if an EGG signal [4, 5] is available. However, this is not usually the case in practice and, therefore, algorithms for automatic GCI identification from the speech signal are preferred. The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) was recently proposed in [6] and was demonstrated to detect accurately the GCIs in anechoic speech recordings.

In many modern telecommunication applications, speech signals are obtained in enclosed spaces such as office rooms, with the talker situated at a distance from the microphone. In this case, the observed speech signal is distorted by reverberation, resulting from sound reflections off the surrounding walls and objects. Reverberation distorts the speech signals [7], acting adversely on many speech processing applications including speech recognition and hands-free telephony. Reverberation will, inevitably, degrade the performance of GCI identification algorithms so it forms an important topic of research for the practical applicability of such algorithms in the future.

In this paper, we first study the effects of reverberation on the performance of the DYPSA algorithm for a single microphone. Next, we propose an extension of DYPSA to the multimicrophone case. Microphone arrays are known to be advantageous for sound capture in reverberant environments [8] due to the spatial diversity of the room transfer function (RTF). In particular we will investigate the application of two different approaches to multimicrophone processing in the context of GCI identification using DYPSA: (i) preprocessing using a delay-and-sum beamformer (DSB), and (ii) implementing an additional penalty function in the dynamic programming (DP) element of DYPSA.

The remainder of this paper is organized as follows. Section 2 presents the effects of reverberation on speech and discusses the consequences on GCI identification. Section 3 reviews the DYPSA algorithm. The extension of DYPSA to the multimicrophone case is presented in Section 4 and supporting simulation results are provided in Section 5. Finally, conclusions are drawn from this work in Section 6.

2. REVERBERATION EFFECTS ON SPEECH

Consider a speech signal $s(n)$ produced in a reverberant room and observed by an M-element microphone array positioned at a distance from the source. The mth microphone observation is

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n), \quad m = 1, 2, \ldots, M \tag{1}$$

where $\mathbf{h}_m = [h_{m,0}\ h_{m,1}\ \ldots\ h_{m,L-1}]^T$ is the L-tap impulse response of the acoustic channel between the source to the mth microphone, $\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L+1)]^T$ is vector of input samples at time n and ^{T} is the transpose operator. The problem is to identify the GCIs in $s(n)$, using the observations $x_m(n)$.

DYPSA operates on the linear prediction (LP) residual, $e(n)$. The LP residual of clean voiced speech is characterized by a quasi-periodic pulse train representing the speech excitation, and approximately constitutes the instants of glottal closure [9]. In general, GCI identification algorithms attempt to locate these peaks [6], which can prove a difficult task; the pulse-train model of the LP residual is over-simplified and doesn't incorporate the noise-like signal components between the excitation peaks.

It has been demonstrated for reverberant speech that the reverberation mainly affects the LP residual. Studies on the effect of reverberation on voiced speech LP residuals [8, 10] have further shown that the room impulse response results in additional spurious peaks of similar amplitude to the original excitation peaks. These erroneous peaks make it difficult to distinguish the true GCIs as shown in Fig. 1. However, in multiple time-aligned observations from a beamformer, the peaks due to GCIs are correlated, while those due to reverberation are not. This observation has motivated the development of several speech dereverberation algorithms [3, 8, 11] which reduce the effects of reverberation by attenuating such uncorrelated components.

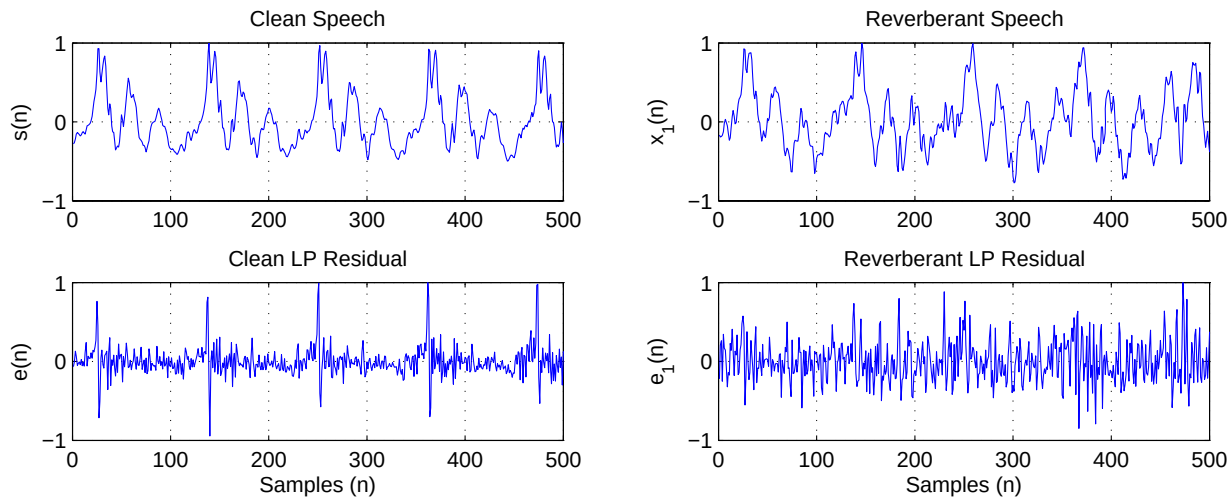

Figure 1: Effects of reverberation on LPC residuals: (left) clean speech and clean residual, (right) reverberant speech and reverberant residual.

3. THE DYPSA ALGORITHM

The main features of the DYPSA algorithm are now reviewed. It consists of three main components: the phase slope function, the phase slope projection, and dynamic programming. These components are defined as follows.

Phase-slope function [12] – defined as the average slope of the unwrapped phase spectrum of the short time Fourier transform of the prediction residual. GCI candidates are selected based on the positive-going zero crossings of the phase-slope function.

Phase-slope projection – introduced to generate GCI candidates when a local minimum is followed by a local maximum without crossing a zero. The midpoint between these is identified and projected onto the time axis with unit slope. In this way, GCIs whose positive going slope does not cross the zero point (those missed by the phase-slope function) are identified.

Dynamic Programming – uses known characteristics of voiced speech and forms a cost function to select a subset of the GCI candidates which are most likely to correspond to the true ones. The subset of candidates is selected according to the minimisation problem defined as

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \lambda^T \mathbf{c}_{\Omega}(r), \tag{2}$$

where Ω is a subset with GCIs of size $|\Omega|$ selected from all GCI candidates, $\lambda = [\lambda_A \ \lambda_P \ \lambda_J \ \lambda_F \ \lambda_S]^T = [0.8 \ 0.5 \ 0.4 \ 0.3 \ 0.1]^T$ is a vector of weighting factors with the values taken here as in [6] and $\mathbf{c}(r) = [c_A(r) \ c_P(r) \ c_J(r) \ c_F(r) \ c_S(r)]^T$ is a vector of cost elements evaluated at the rth GCI of the subset. The cost vector elements are:

- *Speech waveform similarity*, $c_A(r)$, between neighbouring candidates, where candidates not correlated with the previous candidate are penalised.
- *Pitch deviation*, $c_P(r)$, between the current and the previous two candidates, where candidates with large deviation are penalised.
- *Projected candidate cost*, $c_J(r)$, for the candidates from the phase-slope projection, which often arise from erroneous peaks.
- *Normalised energy*, $c_F(r)$, which penalises candidates that do not correspond to high energy in the speech signal.
- *Ideal phase-slope function deviation*, $c_S(r)$, where candidates arising from zero-crossings with gradients close to unity are favoured.

Using the characteristics of the prediction residuals resulting from clean and reverberant speech discussed in Section 2 and the properties of DYPSA presented above, the following remarks can be made:

(i) The reverberant prediction residual contains many peaks due to the room impulse response, whose amplitudes are comparable to the desired peaks in the clean speech residual. Consequently, the phase-slope function and the phase-slope projection are likely to produce many erroneous candidates.

(ii) Peaks of similar amplitude to the true excitation peaks from the clean prediction residual are likely to result in wrong candidates if they both occur in the same analysis frame for the short time Fourier transform.

(iii) A voiced speech segment of weak energy which is preceded by a high energy component is likely to result in erroneous candidates due to the smearing effect of the room impulse response. Such segments occur, for example, at the end of voiced utterances.

It can be seen from the dynamic programming criteria that DYPSA is robust to spurious peaks in the prediction residual. This is an attractive feature for GCI identification in reverberant speech and can be expected to discriminate many of the erroneous candidates due to reverberation. Nevertheless, the performance of DYPSA is degraded significantly with increased reverberation, as will be shown by the simulation results in Section 5. Due to the spatial diversity of the room impulse responses [7], the adverse effects outlined above can be reduced by using multiple microphones which is the motivation for the introduction of multichannel processing within DYPSA.

4. MULTICHANNEL DYPSA

This section presents two approaches to multichannel DYPSA.

4.1 DYPSA at the output of a beamformer

The output of the DSB can be written

$$\bar{x}(n) = \frac{1}{M}\sum_{m=0}^{M-1} x_m(n-\tau_m), \tag{3}$$

where τ_m is a delay to compensate for the time delay of arrival to the different microphones in the array and is assumed to be known. $\bar{x}(n)$ is then presented as a single-channel input to the standard DYPSA algorithm. We refer to this approach as DSB-DYPSA.

4.2 Multichannel Candidate Generation and Selection

Multichannel DYPSA (MC-DYPSA) is a novel extension to DYPSA which relies on the correlation of GCI candidates across multiple channels. As described in Section 3, single-channel DYPSA can be split into three stages: (i) Candidate GCIs are determined by the zero crossings of the phase slope function of the LPC residual, (ii) Points of inflexion which do not cross zero are projected onto the time axis with unit slope to add further candidates, (iii) Dynamic Programming (DP) selects the most likely GCIs based upon a defined cost function. MC-DYPSA performs stages (i) and (ii) on each channel independently. An additional component is incorporated into the DP cost function, which penalizes candidates that are not well correlated across time-aligned channels.

We denote channel $m = \{0,1,\ldots,M-1\}$ containing N samples indexed $n = \{0,1,\ldots,N-1\}$. Each channel contains R_m GCI candidates, enumerated by $r = \{0,1,\ldots,R_m-1\}$, located at samples $n_{r,m}$. Unique GCI candidates (those occurring in at least one channel at the same time) are defined as $n_r = \{n_{r,0} \cup n_{r,1} \cup \ldots \cup n_{r,M-1}\}$, so that n_r is the union of the unique GCI candidate sets from all channels.

Let $g_m(n)$ be a train of impulses at times corresponding to the locations of GCI candidates for channel m, such that

$$g_m(n) = \begin{cases} 1 & n = n_{r,m} \forall r \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The mean, $\bar{g}(n)$, of $g_m(n)$ across all channels is a function indicating the number of occurrences of GCI candidates for a given sample n,

$$\bar{g}(n) = \frac{1}{M}\sum_{m=0}^{M-1}\sum_{r=0}^{R_m-1} \delta(n-n_{r,m}) \tag{5}$$

where $\delta(n-n_{r,m})$ is a unit impulse function with origin at the candidate r in channel m. Small timing errors can occur in the GCI candidates because of poor channel alignment, phase-slope projection errors and sampling noise (at low sampling frequencies). Therefore a spreading function is applied to $\bar{g}(n)$ so that GCI candidates in close proximity incur a lower cost than those spread further apart. A clipped Gaussian was found to be a suitable spreading function, as shown in Fig. 2, denoted by $\Upsilon(n)$,

$$\Upsilon(n) = \begin{cases} ku(n) & 0 \le |ku(n)| \le 1 \\ 1 & |ku(n)| > 1. \end{cases} \tag{6}$$

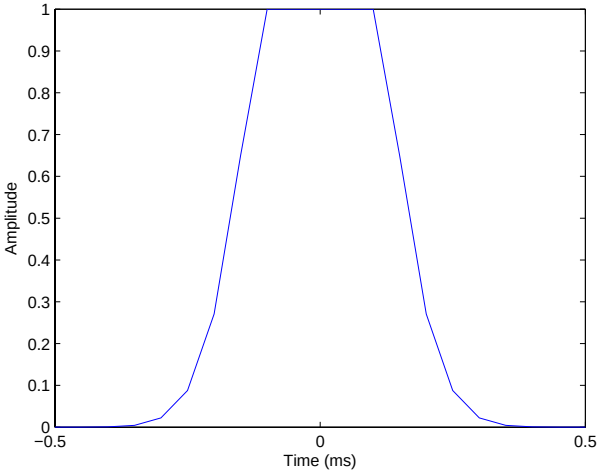

Figure 2: Smoothing Function – a clipped Gaussian for smoothing the cost component for interchannel correlation.

where $u(n)$ is a zero mean unit variance Gaussian multiplied by a gain k. It is convolved with $\bar{g}(n)$ to form a new function $d(n)$,

$$d(n) = \bar{g}(n) * \Upsilon(n) = \frac{1}{M}\sum_{m=0}^{M-1}\sum_{r=0}^{R-1} \delta(n-n_{r,m}) * \Upsilon(n) \tag{7}$$

where $*$ denotes linear convolution. The function $d(n)$ is not bounded in the range $0 < d(n) < 1$ but may exceed 1 depending upon the proximity and height of the samples of $\bar{g}(n)$. Samples for which $d(n)$ exceed 1 are all likely candidates. We next define the inter-channel cost function, $c_I(r)$, such that values of $d(n)$ exceeding 1 are mapped to -0.5 and those in the range $0 < d(n) < 1$ are mapped to $0.5 > d(n) > -0.5$.

$$c_I(r) = \begin{cases} 0.5 - d(n_r) & d(n) < 1 \\ -0.5 & d(n) > 1 \end{cases} \tag{8}$$

Note that this cost function is now a function of r and not n for compatibility with the DYPSA DP. This is a linear mapping for $d(n) < 1$, but it is possible a nonlinear mapping may yield better results by penalising low inter-channel correlation and encouraging high inter-channel correlation to a greater degree.

It was found that the interchannel correlation cost weighting, λ_I, gave best results when set to 0.4.

5. RESULTS

The value T_{60} is defined as the time for a Room Impulse Response (RIR) to decay to -60dB of its initial value. A room measuring 3x4x5 m and T_{60} ranging {100, 150,...,500} ms was simulated using the source-image method [13], containing an array of eight microphones, spaced 50 mm apart, placed on a circular arc 2.5 m from the source so that each channel contained a 2.5 m propagation delay and no inter-channel delay (Fig. 3). Good signal alignment is important and generally requires subsample delays; placing microphones on a circular arc centered at the source alleviates the problem for the purpose of this study. The APLAWD database [14] contains EGG and audio recordings of ten repetitions of five phonetically-balanced English sentences spoken by five male and five female talkers, sampled at 20 kHz.

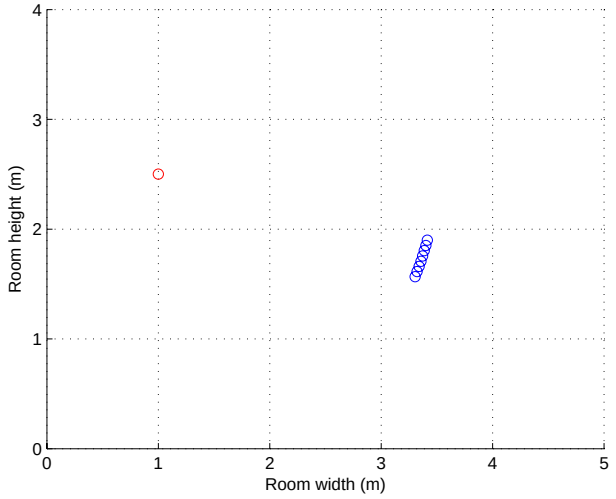

Figure 3: Source and Microphone arrangement. The microphone array is 2.5 m from the source on a circular arc to prevent interchannel delay, removing the necessity for time alignment. The array was placed at a slight angle relative to the walls to reduce strong initial reflections.

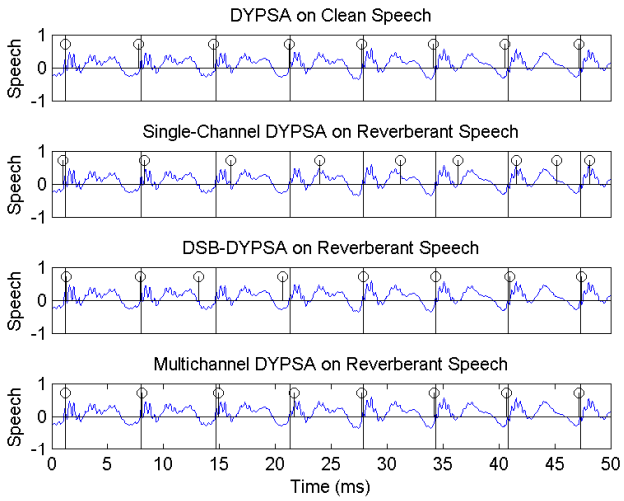

Figure 4: Identified GCIs superimposed onto a clean speech signal for a) DYPSA on clean speech. b) DYPSA on reverberant speech. c) DSB-DYPSA on reverberant speech. d) MC-DYPSA on reverberant speech. Reference GCIs obtained with HQTx are represented by solid vertical lines and estimated GCIs are lines ending in a circle.

The EGG signals were analysed with HQTx [15] to provide reference GCIs. The 19-sample propagation delay from talker to microphone was removed to align the reference and estimated GCIs.

As defined in [6], *Detection rate* is the percentage of all reference GCI periods for which exactly one GCI is estimated. *Accuracy* is the standard deviation of the error between estimated and reference GCIs, when exactly one GCI is estimated in a reference GCI period. *False alarm rate* is the percentage of all reference GCI periods for which more than one GCI is estimated and *Miss rate* is the percentage of all reference GCI periods for which no GCIs were estimated.

5.1 Experiment 1

A speech file from the APLAWD database was analysed with DYPSA. The sample was then convolved with channel 1 of the microphone array in the T_{60}=500 ms case then analysed with DYPSA, DSB-DYPSA and MC-DYPSA. The results depicted in Fig. 4 show eight reference GCIs derived from the associated EGG signal with HQTx as solid vertical lines and estimated GCIs as short lines terminating in a circle, against the clean speech waveform. DYPSA correctly identifies GCIs with small margins of error when operated on clean speech, but accuracy falls and spurious GCIs increase in the reverberant case. Beamformed DYPSA shows improvement with no spurious GCIs but accuracy is significantly lower than the clean case. MC-DYPSA achieves identification on a par with clean DYPSA. This experiment is somewhat idealized which merely demonstrates common errors made by DYPSA and DSB-DYPSA with reverberant speech. MC-DYPSA operating on reverberant speech will not always identify GCIs as well as DYPSA on clean speech.

5.2 Experiment 2

The APLAWD database was convolved with each RIR in turn and analysed with DYPSA, DSB-DYPSA and MC-DYPSA. The results are shown in Fig. 5, Fig. 6 and Table 1. In all cases, the greatest degradation in detection rate occurs in the lower increments of T_{60} and tails off gently with high reverberation. Single-channel DYPSA shows the worst degradation, dropping by 8% between clean and T_{60}=100 ms and 31% at T_{60}=500 ms. Multichannel achieves the best with a 12% drop at T_{60}=500 ms. Miss and false alarm rates also show significant improvement.

Like detection rate, the greatest degradation in accuracy occurs in the first few increments of T_{60} and tails off with higher reverberation. MC-DYPSA has a higher hit rate so more candidates are included in the calculation of accuracy, causing MC-DYPSA to appear to degrade further than DYPSA and DSB-DYPSA with high T_{60}. Note that hit rate and accuracy from clean DYPSA differ slightly to those given in [6] because the reference GCIs were derived from a newer version of HQTx.

6. CONCLUSIONS

The DYPSA algorithm is a robust method for GCI extraction from voiced speech with low levels of reverberation. However, recording environments such as offices often cause significant sound reflection, resulting in reverberation and limiting the applicability of DYPSA in these situations. A microphone array and DSB used as a preprocessor to DYPSA can significantly improve the estimation of GCIs and may provide acceptable results in environments with moderate levels of reverberation. Multichannel DYPSA is an extension to DYPSA which uses the correlation of GCI candidates from each microphone in an array to provide highly robust GCI estimation. Though MC-DYPSA contains many parameters which require optimization, preliminary results presented in this paper suggest that the adopted approach yields very good GCI estimation in highly reverberant environments.

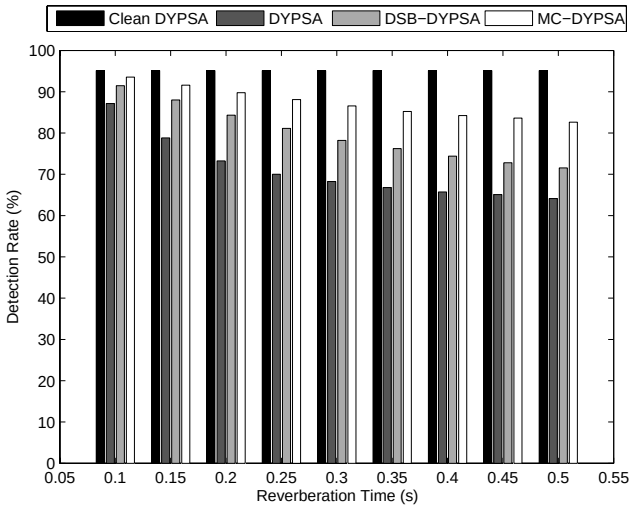

Figure 5: Detection rate vs. reverberation time for DYPSA on clean speech, DYPSA on reverberant speech, DSB-DYPSA on reverberant speech and MC-DYPSA on reverberant speech.

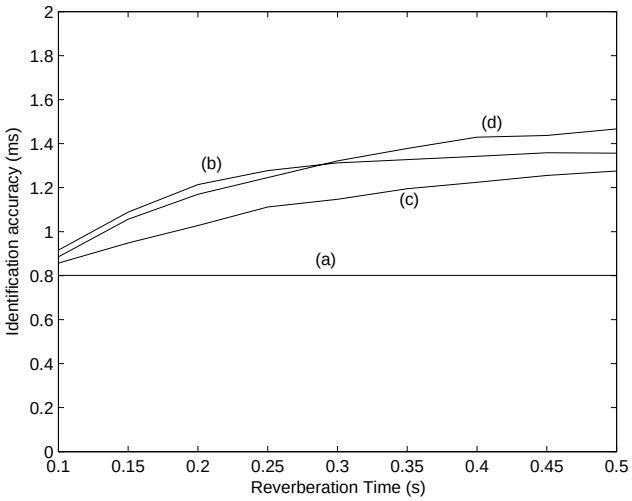

Figure 6: Identification accuracy vs. reverberation time for a) DYPSA on clean speech, b) DYPSA on reverberant speech, c) DSB-DYPSA on reverberant speech, d) MC-DYPSA on reverberant speech.

Table 1: Performance comparison for DYPSA algorithms on the APLAWD database.

	ID Rate (%)	Miss Rate (%)	FA Rate (%)	ID Acc., σ (ms)
Clean DYPSA	95.1	2.3	2.6	0.80
0.1s DYPSA	87.1	4.1	8.8	0.92
0.1s DSB-DYPSA	91.5	3.3	5.3	0.86
0.1s MC-DYPSA	93.5	2.5	4.0	0.89
0.5s DYPSA	64.1	7.4	28.5	1.36
0.5s DSB-DYPSA	71.5	6.6	21.8	1.27
0.5s MC-DYPSA	82.6	4.1	13.3	1.46

REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, Dec. 1990.

[2] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Commun.*, vol. 11, no. 2, pp. 175–187, June 1992.

[3] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Mulitmicrophone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 809–812.

[4] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no.8, pp. 730–743, 1986.

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

[6] 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, Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[7] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Francis, Oct. 2000.

[8] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Springer-Verlag, Berlin, 2001.

[9] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*. Macmillan, 1993.

[10] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[11] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2002, pp. 541–544.

[12] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Processing*, vol. 3 no. 9, pp. 325–333, 1995.

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

[14] G. Lindsey, A. Breen, and S. Nevard, "Archivable actual-word databases," Univ. College London, London, U.K., Tech. Rep., Jun. 1987.

[15] M. Huckvale, "Speech filing system: Tools for speech," Research University College London [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs, Tech. Rep., 2004.