ACOUSTIC FEEDBACK CANCELLATION FOR LONG ACOUSTIC PATHS USING A NONSTATIONARY SOURCE MODEL

G. Rombouts, T. van Waterschoot, K. Struyve(*), M. Moonen

Katholieke Universiteit Leuven, ESAT-SCD
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{rombouts,tvanwate,moonen}@esat.kuleuven.ac.be
(*) Televic N.V., L. Bekaertlaan 1, 8870 Izegem, Belgium
k.struyve@televic.com

ABSTRACT

Several pro–active acoustic feedback (Larsen–effect) cancellation schemes have been presented for speech applications with short acoustic feedback paths as encountered in hearing aids, but these schemes fail with the long impulse responses inherent to public address systems. We derive a new prediction error method (PEM) based scheme (referred to as PEM–AFROW) which identifies both the acoustic feedback path and the nonstationary speech source model. A cascade of a short– and a long term predictor removes the coloring and periodicity in voiced speech segments, which account for the unwanted correlation between the loudspeaker signal and the speech source signal. The predictors calculate row operations which are applied to pre–whiten a least squares system, which is then solved recursively by means of e.g. NLMS or RLS algorithms. Simulations show that this approach is indeed superior to earlier approaches whenever long acoustic channels are dealt with.

1. INTRODUCTION

Acoustic feedback, also referred to as the *Larsen–effect* (howling) occurs in microphone–amplifier–loudspeaker–room systems when the loop gain is larger than one at a frequency where the loop phase is a multiple of 2π.

A conventional solution consists of inserting notch filters into the signal path, thus decreasing the loop gain at those frequencies for which the problem arises. There are several disadvantages to this approach : the system is reactive (the howling phenomenon occurs for about 0.5 seconds before it is detected), the desired signal is distorted by the notch filters, and the 'reverb–like' sound which occurs in a system which is marginally stable is not suppressed.

In this paper, we will focus on single channel acoustic feedback cancellation (AFC) schemes as depicted in **Figure 1**. This setup does not exhibit the disadvantages summarized above. The estimate of the filter coefficient vector $\mathbf{f}(k)$ of the acoustic path $F(q,k) = \mathbf{f}(k)^T\mathbf{q} = f_0q^0 + \ldots + f_{N-1}q^{-N+1}$ from the loudspeaker to the microphone is $\hat{\mathbf{f}}(k)$. Here q^{-1} is the delay operator. The N coefficients of $\hat{\mathbf{f}}(k)$ are copied at regular time instants to the cancellation filter $\hat{\mathbf{f}}_0(k)$. The loudspeaker signal $u(k)$ is filtered by the room impulse response $\mathbf{f}(k)$ and also by the cancellation filter $\hat{\mathbf{f}}_0(k)$. The difference between the cancellation filter output and the microphone signal is the error signal $e(k)$ which should then be equal to the speech source signal $v(k)$ (for a correct model $\hat{\mathbf{f}}(k)$). In **Figure 1**, g is the amplifier gain, $y(k)$ is the microphone signal, $u(k) = ge(k)$ is the loudspeaker signal, $\mathbf{f}$ is the feedback path impulse response, $v(k)$ is the (speech) source signal, $w(k)$ is the excitation sequence of the source signal, and $H(q,k) = (1 + a_1(k)q^{-1} + \ldots + a_P(k)q^{-P})^{-1}$ is a time varying autoregressive (AR) speech model of order P. The coefficient vector of the numerator is $\mathbf{a}(k)$. Finally, the q^{-D} block in **Figure 1** is a forward delay, which is often unavoidable in digital implementations (buffers for AD/DA-converters, ...) , but which will be exploited further on.

An acoustic echo cancellation (AEC) like approach has been used for AFC in e.g. [1, 2]. The main complication in AFC compared to the direct identification-approach used in AEC is that in AFC, one can not assume that the speech source signal $v(k)$ is uncorrelated with the loudspeaker signal $u(k)$. Ignoring this, and applying direct identification anyway, would result in a bias in the identified room impulse response [3, 4]. This bias can be removed using the prediction error method (PEM), which incorporates a speech source signal model into the identification procedure [5]. This has been studied mainly in the hearing aids context where the feedback path impulse response is less than 5 msec. In this paper, we focus on public address (PA) systems, where the feedback path typically has a much longer impulse response, e.g. up to 500 msec, and hence an alternative approach will be needed.

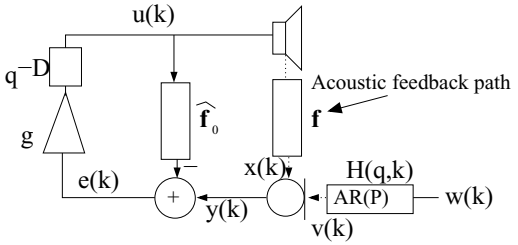

Figure 1: Acoustic feedback cancellation scheme

Speech, although highly nonstationary over longer time periods, is often considered to be stationary during short frames of ca. 20 msec (e.g. 160 samples at 8 kHz). Within these frames, it can be whitened by a cascade of a short term predictor (STP) and a long term predictor (LTP). It is required to use data windows of several seconds to estimate the room impulse response, over the length of which the speech signal will be nonstationary. This contrast between the long stationarity period of the long room impulse response and the short stationarity period of the short term prediction speech model, and the corresponding number of data points which are available to identify each of them, is fundamental to the problem of acoustic feedback cancellation for public address systems.

In this paper, we introduce a new technique which estimates the speech model over short time windows (over which it is stationary), and the room impulse model over longer time windows (which is necessary because the number of parameters is much larger). The speech model is not required to be stationary during the complete room impulse response. Our scheme will also include a long time predictor which models the periodicity in $w(k)$. We will show that this scheme outperforms existing methods.

This paper is organized as follows. In section 2, we introduce our new procedure. It uses alternating updates of the speech model and the adaptive filter which models the room. An important difference with [5] is that in our algorithm the speech model provides row transformations, which are then applied to pre–whiten the least

squares system from which the room response estimate is computed. Hence the name 'prediction error method based adaptive filtering with row operations' (PEM-AFROW). In section 3, complexity figures are given, in section 4 we show simulation results, and section 5 contains the conclusion of the paper.

2. PEM-AFROW

It is instructive to first assume that $w(k)$, the excitation sequence, is a white noise sequence. This means that we model the speech source signal as a time varying AR (TVAR) signal of order P, i.e. $v(k) = a_1 v(k-1) + ... + a_P v(k-P) + b(k)w(k)$. Here $b(k)$ accounts for energy variations in the excitation signal. Later on, we will use a more general model for $w(k)$. We start from the minimization problem

$$\min_{\hat{\mathbf{f}}} \left\| U(k)\hat{\mathbf{f}}(k) - \mathbf{y}(k) \right\|, \tag{1}$$

with

$$U(k) = \begin{pmatrix} u(k) & u(k-1) & \cdots & u(k-N+1) \\ u(k-1) & & & \\ \vdots & \vdots & \ddots & \vdots \\ u(0) & 0 & \cdots & 0 \end{pmatrix}.$$

and $\mathbf{y}(k) = (\ y(k) \quad \ldots \quad y(0)\ )^T$. This minimization results in a biased estimate $\hat{\mathbf{f}}(k)$ for $\mathbf{f}(k)$ since $y(k)$ contains the 'disturbance' $v(k)$ which is correlated with $u(k)$: $\mathbf{y}(k) = U(k)\mathbf{f}(k) + \mathbf{v}(k)$. If we have an estimate $\hat{A}(q,k)$ of $H^{-1}(q,k)$ available at each time instant, with coefficients $\hat{\mathbf{a}}(k) \in \mathbb{R}^P$, we can apply a pre-whitening by forming the matrix

$$\hat{A}(k) = \begin{pmatrix} \hat{\mathbf{a}}^T(k) & 0 & 0 & 0 \\ 0 & \hat{\mathbf{a}}^T(k-1) & 0 & 0 \\ 0 & 0 & \hat{\mathbf{a}}^T(k-2) & 0 \\ 0 & 0 & 0 & \ddots \end{pmatrix}.$$

It is important to note that each row in the matrix is shifted over one position compared to the previous row, hence that the second row has one zero in front of the transposed vector $\hat{\mathbf{a}}^T(k-1)$ of dimension $P+1$, the third row has two zeros, ... We can now modify the minimisation problem (1) to

$$\min_{\hat{\mathbf{f}}} \left\| \hat{A}(k)U(k)\hat{\mathbf{f}}(k) - \hat{A}(k)\mathbf{y}(k) \right\|. \tag{2}$$

We now introduce the assumption that $\mathbf{h}(k)$ is constant during frames of 20 msec. This means that we rewrite

$$\hat{A}(k) = \begin{pmatrix} \hat{\mathbf{a}}_i^T & 0 & 0 & 0 \\ \ddots & & & \\ 0 & \hat{\mathbf{a}}_i^T & 0 & 0 \\ 0 & 0 & \hat{\mathbf{a}}_{i-1}^T & 0 \\ 0 & 0 & 0 & \ddots \end{pmatrix} \begin{matrix} \uparrow \\ L \\ \downarrow \\ \uparrow \\ L \\ \downarrow \end{matrix}, \tag{3}$$

with e.g. $L = 160$ for a sampling rate of 8 kHz, and $i = \lceil k/L \rceil$, the first integer larger than k/L. This means that i is the frame index.

We now decouple the non–linear equations in order to also calculate the estimates $\hat{\mathbf{a}}_i$ and the room impulse response $\hat{\mathbf{f}}(k)$ in an alternating fashion. In the first step, a previous estimate $\hat{\mathbf{f}}(k)$ is used to filter a frame of data (20 msec). The filter output is subtracted from the corresponding microphone samples, resulting in $d(k) = y(k) - U(k)\hat{\mathbf{f}}(k)$. Note that if the estimate $\hat{\mathbf{f}}(k)$ were exact, then $d(k) = v(k)$.

Linear prediction is then performed on this $d(k)$ (Levinson-Durbin) algorithm to find the linear prediction error filter $\hat{\mathbf{a}}_i$. In the second step, (2) is solved for $\hat{\mathbf{f}}(k)$ with the updated (fixed) value for $\hat{\mathbf{a}}_i$. This gives a better estimate $\hat{\mathbf{f}}(k)$ for $\mathbf{f}(k)$. These two steps can be iterated on the frame. Since none of these two steps will increase $\varepsilon\{\|\mathbf{e}(k)\|\} = \varepsilon\{\left\|\hat{A}(k)U(k)\hat{\mathbf{f}}(k) - \hat{A}(k)y(k)\right\|\}$, the algorithm will converge to a (possibly local) minimum of (2).

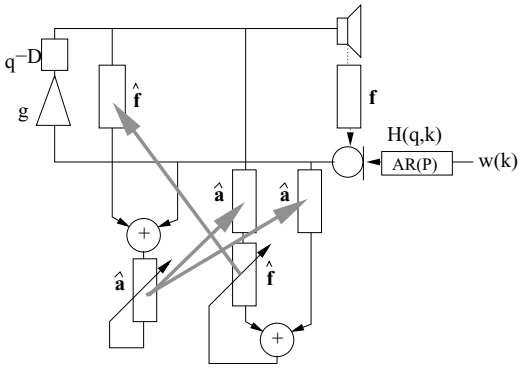

Figure 2: PEM-AFROW-identification. In the first phase, $\hat{\mathbf{a}}$ is estimated in the left hand side, it is then copied to the right hand side, where the estimation of $\hat{\mathbf{f}}$ is performed on the same data frame. Finally, $\hat{\mathbf{f}}$ is copied to the left hand side and used in the next frame.

In order to reduce the complexity, we will perform only one iteration per frame. The minimization problem (2), with a fixed value of $\hat{A}(k)$, can be solved for $\hat{\mathbf{f}}(k)$ by means of any adaptive filtering algorithm. We have implemented this both using a QRD–based RLS algorithm and an NLMS algorithm. The input vector is in both cases

$$\mathbf{u}_w^T(k) = \mathbf{a}_i^T \begin{pmatrix} \mathbf{u}^T(k) \\ \vdots \\ \mathbf{u}^T(k-P+1) \end{pmatrix}, i = \lceil k/L \rceil, \tag{4}$$

Here $\mathbf{u}(k) = (\ u(k) \quad \ldots \quad u(k-N+1)\ )^T$. The desired signal input (right hand side sample) is $\mathbf{a}_i^T\mathbf{y}(k)$ with $\mathbf{y}(k) = (\ y(k) \quad \ldots \quad y(k-P+1)\ )^T$. Since $\mathbf{u}(k)$ is a shifted version of $\mathbf{u}(k-1)$ with one sample prepended, and $\mathbf{a}_i$ remains constant during a frame of L samples, $\mathbf{u}_w(k)$ will be a shifted version of $\mathbf{u}_w(k-1)$ with one sample prepended. So inside a frame, only one vector multiplication has to be performed to calculate $\mathbf{u}_w(k)$. On the other hand, at the start of each frame, a matrix multiplication should be performed to calculate all elements of $\mathbf{u}_w(iL)$ as follows :

$$\mathbf{u}_w^T((i-1)L+1) = \mathbf{a}_i^T \begin{pmatrix} \mathbf{u}^T((i-1)L+1) \\ \vdots \\ \mathbf{u}^T((i-1)L-P+2) \end{pmatrix}.$$

The identification algorithm is shown in **Figure** 2. For real time implementation, the scheme involves a delay of one frame for the update of $\hat{\mathbf{f}}(k)$, since $\mathbf{a}_i$ can only be calculated at time iL. Note that this is not a problem since we have assumed that the room impulse response is constant over more than one frame. The delay is effectively implemented as a delay line for the input samples $u(k)$ before they are fed to equation (4).

Once the room impulse response has been identified, the next step is to insert the cancellation filter into the feedback loop scheme by setting $\hat{\mathbf{f}}_0(k) = \hat{\mathbf{f}}(k)$, e.g. at regular time intervals (see **Figure** 1). It is important to notice that this obviously influences the adaptation. The input data used for the identification procedure then depend on the current model estimate, which is reminiscent of a non–linear optimization problem. This dependency is effectively ignored in

our implementation (it is also ignored in adaptive control theory [6]).

Experiments indicate that updating the cancellation filter regularly is beneficial to the identification process. This can be explained because a time variant forward path (from microphone to loudspeaker) decreases the correlation between the loudspeaker signal and the speech source signal.

At this point, the difference between PEM–AFC and PEM-AFROW becomes obvious : in PEM–AFROW the stationarity of the speech model is explicitly assumed in the minimization problem by stating that $\hat{\mathbf{a}}_i$ remains constant during a frame (see equation (3)). At the start of each frame, the full input vector $\mathbf{u}_w^T(k)$ is recalculated. In PEM–AFC, this assumption of stationarity is not made for the optimisation problem itself (the optimisation is decoupled in two completely independent adaptive filters), and the full input vector is never recomputed after a change of $\mathbf{a}(k)$ in PEM–AFC, which can only be justified for short impulse responses.

For the TVAR-signals we studied up till now (where $w(k)$ was a white noise sequence) , the pre–whitening step removes all of the correlation between the loudspeaker signal and the source signal. However, the excitation sequence $w(k)$ for voiced speech is periodical (glottal excitation). Hence the input signal $u(k)$ of the adaptive filter is — due to this periodicity — still correlated with the source signal, even after pre–whitening.

A standard approach in speech coding [7] is to cascade a short term predictor (STP) of order P (e.g. 12) which models the vocal tract characteristics,

$$\mathbf{u}_{sw}^T(k) = \mathbf{a}_i^T \begin{pmatrix} \mathbf{u}^T(k) \\ \vdots \\ \mathbf{u}^T(k-P+1) \end{pmatrix}, i = \lceil k/L \rceil , \qquad (5)$$

with a long term predictor (LTP) with only one tap and a lag equal to the pitch period to model the periodicity, $u_{lsw}(k) = u_{sw}(k) + b_j u_{sw}(k-M_j), j = \lceil k/L_{ltp} \rceil$ .The LTP can be estimated in windows of 20 msec (which is the frame length L of the short term predictor), with a 10 msec overlap. This means that the LTP model is estimated each 10 msec, which corresponds to L_{ltp} samples (at 8 kHz, $L_{ltp} = 80$). In order to estimate the LTP, we minimize $E_j == \min \varepsilon\{\left\|u_{sw}(k-M_j)b_j + u_{sw}(k)\right\|^2\}$. The solution follows from $\varepsilon\{u_{sw}(k-M_j)u_{sw}(k-M_j)\}.b_j = \varepsilon\{u_{sw}(k)u_{sw}(k-M_j)\}$. We can now estimate the one long term prediction filter tap $b_j = (\overline{\mathbf{u}}_{sw}^T(k-M_j)\overline{\mathbf{u}}_{sw}(k-M_j))^{-1}\overline{\mathbf{u}}_{sw}^T(k)\overline{\mathbf{u}}_{sw}(k-M_j)$. In this equation $\overline{\mathbf{u}}_{sw}(k) = (\ u_{sw}(k) \quad \ldots \quad u_{sw}(k-L+1)\ )^T$. The variance of the long term prediction residual is

$$E_j = \overline{\mathbf{u}}_{sw}^T(k)\overline{\mathbf{u}}_{sw}(k) - \frac{(\overline{\mathbf{u}}_{sw}^T(k)\overline{\mathbf{u}}_{sw}(k-M_j))^2}{\overline{\mathbf{u}}_{sw}^T(k-M_j)\overline{\mathbf{u}}_{sw}(k-M_j)}$$

This is evaluated for different values of $M_{j,i} = M_{\min} \ldots M_{\max}$ (the lag), and the parameters (M_j, b_j) which result in the minimum value of E_j are chosen as the predictor for long term prediction frame j.

It is important to note that by applying long term prediction, the actual order of the speech source model is the lag of the long term model plus the order of the short term model, and as stated in [5], to guarantee identifiability, the forward delay must be larger than the order of this model. In practice it does not matter too much where this forward delay is implemented : often a latency D is introduced by buffering after and before the A/D and D/A–converters, or even — due to the relatively low velocity of sound waves — from the distance between the loudspeaker and the microphone.

In section 2 it was mentioned that at frame borders, the whole input vector has to be recalculated by means of a matrix multiplication. It must be noted that when long term prediction is added to the algorithm, this matrix multiplication has to be performed not only at frame borders of the short term predictor, but also at frame borders of the long term predictor.

3. COMPLEXITY

The complexity is evaluated when the algorithm is operated with an NLMS adaptive filter. In these complexity expressions a multiplication and an addition are counted as two separate floating point operations. A 'search range' $M_{\min}$ to $M_{\max}$ has to be specified for the lag of the long term predictor (typically $M_{\min} = 20$, $M_{\max} = 160$ at 8 kHz). The complexity depends on these parameters through $dM = M_{\max} - M_{\min}$. For the complexity calculation we assume one tap long term prediction, and we also assume that the frames do not overlap. Since at each frame border the full NLMS input vector is recalculated, the complexity per sample is $8(N+P)+4dM+5+((2P+4)N+4P^2-5P+15)/L$ floating point operations. The algorithm was implemented in C++ on a Pentium III, 1GHz PC without any specific optimization effort, and runs in real time with $N = 2000$, $P = 12$, $L = 160$ at 16 kHz sampling rate, with long term prediction overlap of 80 samples. In case of no overlap for the long term predictor, the number of floating point operations per second would be 272.10^6.

4. SIMULATION RESULTS

In **Figure** 3 the error norm$\left\|\mathbf{f}(k) - \hat{\mathbf{f}}(k)\right\|$is plotted as a function of time. Note that only the identification performance is shown, which means that the cancellation filter is *not* inserted into the scheme during adaptation. The signal is a sentence uttered by a male voice, the acoustic path has 1000 taps. We use NLMS for the adaptive filter, since in a practical implementation this would be the adaptive algorithm of choice (due to complexity constraints). Note that the performance of all algorithms is dependent of the energy ratio ('signal to noise ratio') of the loudspeaker component arriving on the microphone versus the source signal arriving on the microphone (the source signal should thus be interpreted as 'noise'). The simulations shown here were done for one specific situation where this ratio was -11 dB, but experiments show a similar performance difference between the algorithms for other ratios. The short time prediction frame length is 160 samples, the long term prediction frame overlap is 80 samples, the minimum– and maximum lag for the long term predictor are 20 and 160 respectively. the sampling frequency is 8 kHz. The speech model order in PEM–AFROW and PEM–AFC is 12. The forward delay is 200 taps in both PEM–AFROW and PEM–AFC (note that the PEM–AFC version of [8] does not explicitly incorporate a forward delay, but the theoretical analysis of [5] shows that this is required for correct performance, hence we added it to the system). We also show the performance of PEM-AFROW with the long term predictor disabled, because PEM–AFC also does not use a long term predictor.

The NLMS step size is 0.01 for PEM–AFROW, while PEM–AFC, which uses a modified NLMS algorithm and hence a different definition of the step size, was tuned to give the same initial convergence speed. This allows us to make a fair comparison of the resulting bias/variance of the solution. Direct identification is seen to give poor results. PEM–AFC performance decreases with path length, and for $N = 1000$, its behaviour is only slightly better than direct identification behaviour (i.e. when the room impulse response is identified as if the system were operating in open loop). PEM–AFROW does perform well also for long paths. The bad performance of PEM-AFC is to be attributed to the stationary speech model assumption, which is not fulfilled for long paths.

5. CONCLUSION

We have introduced a new algorithm, referred to as PEM–AFROW, which allows for acoustic feedback cancellation in setups with long acoustic paths. It uses a speech source model with short– and long term prediction. Not only the howling phenomenon is suppressed but also the reverberation–like sounds, which become audible in the marginal stability region. The main differences with existing schemes are that our algorithm incorporates a long term prediction filter which removes periodicity in the short term speech signal residual, and that we do not assume stationarity of the speech signal

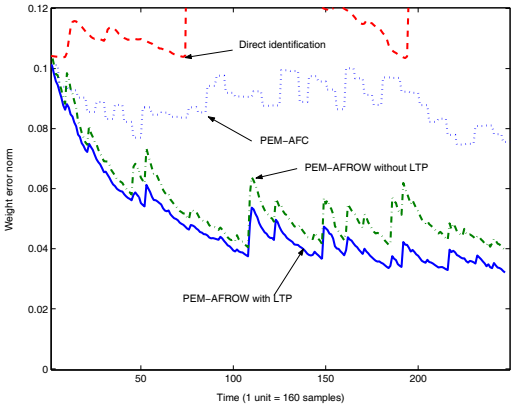

Figure 3: PEM–AFROW with and without long term prediction versus direct identification for long paths (1000 filter taps).

over the length of the data window on which the acoustic path is identified. PEM-AFROW hence performs very well for long acoustic paths, while it is even slightly better than the existing methods for short path applications. Thanks to the low complexity, the algorithm can easily be implemented in real time.

6. ACKNOWLEDGEMENTS

This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, in the frame of IWT project 020476: 'SMS4PA : Sound Management System For Public Address Systems', Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling') , the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) of the Flemish Government , Research Project FWO nr.G.0233.01 ('Signal processing and automatic patient fitting for advanced auditory prostheses'). The scientific responsibility is assumed by its authors.

REFERENCES

[1] S. Kamerling, K. Janse, and F. van der Meulen, "A new way of acoustic feedback suppression," in *AES*. AES, 1998.

[2] C. P. Janse and P. A. A. Timmermans, "Signal amplifier system with improved echo cancellation," U.S. Patent 5,748,751, May 1998.

[3] M. G. Siqueira and A. Alwan, "Steady–state analysis of continuous adaptation in acoustic feedback reduction systems for hearing–aids," *IEEE Trans. Speech and Audio processing*, vol. 8, no. 4, pp. 443–453, July 2000.

[4] J. Hellgren and U. Forssell, "Bias of feedback cancellation algorithms in hearing aids based on direct closed loop identification," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 7, pp. 906–913, Nov. 2001.

[5] A. Spriet, I. Proudler, M. Moonen, and J. Wouters, "Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal," KULeuven/ESAT-SCD Internal report, ftp.esat.kuleuven.ac.be/pub/sista/spriet/reports/03-167.pdf, August 2003, Accepted for publication in IEEE Transactions on Signal Processing.

[6] L. Ljung and S. Soderstrom, *Theory and Practice of Recursive Identification*, chapter 7, MIT Press, 1983.

[7] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Transactions on acoustisc, speech and signal processing*, vol. 37, no. 4, pp. 467–477, April 1989.

[8] R. Leber and W. Schaub, "Circuit and method for the adaptive suppression of an acoustic feedback," U.S. Patent US6611600, 2003.