ABSTRACT

A lattice-ladder adaptive decorrelation filtering (ADF) algorithm is proposed with the aim of developing a more efficient co-channel speech separation system. It is shown that based on the joint forward and backward linear predictions, a lattice-ladder structure can be derived for ADF. Experimental results show that the proposed algorithm is effective in reducing cross-interference between co-channel speech sources and has better tracking ability on the variations in acoustic environment compared to the original direct-form ADF algorithm.

1. INTRODUCTION

The problem of separating co-channel speech signals produced by competing talkers is becoming an active topic of research in the area of speech and signal processing. A number of co-channel speech separation algorithms utilizing multi-microphone acquisition and second-order statistics have been proposed in the literature [1][2] and shown effective in separating speech sources from their convolutive mixtures in a variety of applications. In our previous work [3], a co-channel speech separation system was developed based on the adaptive decorrelation filtering (ADF) algorithm proposed by Weinstein, Feder, and Oppenheim [1]. The separation system significantly improved accuracy on co-channel speech recognition as well as intelligibility on clinical subject listening test [4].

The current ADF algorithm has several shortcomings due to the use of direct FIR form in modeling acoustic paths. First of all, it is often difficult to know the required filter length in advance. Secondly, the impulse responses of acoustic paths are usually sparsely spaced, making the direct FIR implementation inefficient. Finally, a slight change in acoustic environment can cause significant changes in FIR coefficients and it can take a long time for the adaptation algorithm to adjust for such a change.

Lattice-ladder structure has been widely used in adaptive signal processing such as linear prediction and noise cancellation [5]-[9]. Lattice-ladder filters are modular in structure, so that additional stages can be added when necessary without affecting the earlier stages [9]. This provides the potential flexibility to adapt filter length according to environment. Furthermore, when converting a direct-form FIR filter into the lattice-ladder filter, the filter coefficients become more evenly distributed, which enables more efficient filter estimation and adaptation.

In the current work, a lattice-ladder implementation for the ADF algorithm is proposed with the aim of developing a more efficient co-channel speech separation system. This paper is organized into six sections. In Section 2, co-channel speech separation is briefly discussed and the decorrelation filtering problem is defined. In Section 3, the lattice-ladder structure for decorrelation filtering is developed and in Section 4, the lattice-ladder structure is applied to co-channel speech separation. Experimental results are presented in Section 5 and a conclusion is made in Section 6.

2. FUNDAMENTALS

2.1. Co-Channel Speech Separation

When there are two speech sources in a co-channel environment, the acquired signals are usually convolutive mixtures of the source speech signals. By denoting the speech signals as $x_1(t)$ and $x_2(t)$ and the acquired co-channel mixtures as $y_1(t)$ and $y_2(t)$, the co-channel environment can be modeled in the frequency domain as

$$Y_1(f) = H_{11}(f)X_1(f) + H_{12}(f)X_2(f)$$
$$Y_2(f) = H_{21}(f)X_1(f) + H_{22}(f)X_2(f)$$

where $H_{ij}(f)$ represents the transfer function modeling the acoustic paths from the source $j$ to the microphone $i$. In [1], it is shown that by using the following separation system

$$V_1(f) = Y_1(f) - F_{12}(f)Y_2(f)$$
$$V_2(f) = Y_2(f) - F_{21}(f)V_1(f)$$

the source speech signals can be separated if $F_{12} = H_{12}/H_{22}$ and $F_{21} = H_{21}/H_{11}$. Since $H_{ij}$'s are usually unknown and time-varying, $F_{ij}$'s need to be estimated and tracked.

Assuming that $x_i(t)$'s are zero-mean and uncorrelated, the separated signals $v_i(t)$'s should also be zero-mean and uncorrelated. In many situations, the decorrelation between $v_i(t)$ is a simple and effective criterion for the estimation of $F_{ij}$'s. Define the length-$m$ vector of a signal $x(t)$ as $x_M(t) = [x(t) \ldots x(t-m+1)]^T$. If $F_{ij}$'s are chosen to be length-$M$ FIR filters with coefficients

$$f_{ij} = [f_{ij}(0) \ldots f_{ij}(M-1)]^T$$

then for $v_i(t)$'s to be uncorrelated, $f_{ij}$'s need to satisfy

$$E\left\{v_{ij}(t)\right\} = 0$$

2.2. Decorrelation Filtering

Let $y(t)$, $v(t)$, and $d(t)$ be zero-mean random processes and define the output signal $z(t)$ as

$$z(t) = d(t) - y_M^T(t)\hat{h}$$

In order to decorrelate $z(t)$ and $v(t)$, the length-$m$ decorrelation filter $\hat{h}$ needs to satisfy

$$h^* = E\left\{y_M(t)^Ty_M^T(t)\right\}^{-1} E\left\{z_M(t)d(t)\right\}$$

The lattice-ladder structure for this adaptive filter is first developed in Section 3 and then applied to co-channel speech separation in Section 4.

3. LATTICE-LADDER STRUCTURE FOR ADF

In order to derive the lattice-ladder structure for the decorrelation filter, the joint forward and backward linear predictions of signals $y(t)$ and $v(t)$ are first defined. Then the order adaptation equations (OAEs) for various prediction vectors and the decorrelation filter are derived, which establishes the lattice-ladder structure.

The following notations are used throughout this paper:

- $M$: length of the filter
- $F$: filter coefficients
- $y(t)$: input signal
- $z(t)$: output signal
- $d(t)$: desired signal
- $h$: filter coefficients
- $E$: expectation operator
- $\hat{h}$: estimated filter coefficients
1. Reversed vectors:
V_1 = [p_1, \ldots, p_m]^T and \_2 = [q_1, \ldots, q_m]^T
are reversed vectors if p_i = q_{m-i+1} for i = 1, \ldots, m.
This relation is denoted as p = q^T.

2. Correlation function:
\rho_{x_1 x_2}(\tau) = E \{x_1(t)x_2(t-\tau)\}

3. Correlation matrices of order m:
\begin{align*}
P_m &= E \{y_m(t)p_m^T(t)\} \\
Q_m &= E \{y_m(t)q_m^T(t)\} = P_m^T
\end{align*}
Both of them are Toeplitz matrices consisting of \rho_{y_m}(\tau), \tau = -m+1, \ldots, m-1.

4. Correlation vectors of order m:
\begin{align*}
P_{f,m} &= E \{y_m(t-1)y(t)\} = [\rho_{y_1}(1) \ldots \rho_{y_m}(m)]^T \\
P_{b,m} &= E \{y_m(t)y(t-m)\} = [\rho_{y_1}(-m) \ldots \rho_{y_m}(-1)]^T \\
q_{f,m} &= E \{y_m(t-1)v(t)\} = P_{b,m}^{-1} \\
q_{b,m} &= E \{q_m(t)v(t-m)\} = E^T \\
\rho_m &= E \{y_m(t)v(t)\} = [\rho_{dv}(0) \ldots \rho_{dv}(m-1)]^T
\end{align*}

3.1. Joint Linear Predictions
Traditional lattice-ladder filters are based on forward and backward linear predictions of the input signal. The lattice-ladder decorrelation filter requires similar building blocks: the joint linear predictions of the input signals y(t) and v(t).

Denote the m-th order forward prediction vectors for signals y(t) and v(t) as \_f,m and \_b,m respectively. The m-th order forward prediction errors for signals y(t) and v(t) can be correspondingly defined as
\begin{align*}
e_{pf,m}(t) &= y(t) - y_{pf}(t-1)f \\
e_{qf,m}(t) &= v(t) - y_{qf}(t-1)b
\end{align*}
In order to achieve the solution in Eq. (5), it is necessary for the joint forward linear prediction to find the forward prediction vectors
\begin{align*}
P^*_{f,m} = P_m^{-1}P_{f,m} \\
q^*_{f,m} = Q_m^{-1}q_{f,m}
\end{align*}
such that the resulting cross-correlation between the forward prediction errors \_pf,m(t) and \_qf,m(t) becomes
\begin{align*}
E_{pf,m} &= r_{ys}(\theta) - q^T_{f,m}P_m^*P_{f,m}
\end{align*}

Similarly denote the m-th order backward prediction vectors for signals y(t) and v(t) as \_b,p,m and \_b,q,m respectively. The m-th order backward prediction errors for signals y(t) and v(t) can be correspondingly defined as
\begin{align*}
e_{pb,m}(t) &= y(t-m) - y_{pb}(t-1)b \\
e_{qb,m}(t) &= v(t-m) - y_{qb}(t-1)b
\end{align*}
In the joint forward linear prediction, the goal here is to find the backward prediction vectors
\begin{align*}
P^*_{b,m} = P_m^*P_{b,m} \\
q^*_{b,m} = Q_m^*q_{b,m}
\end{align*}
such that the resulting cross-correlation between the backward prediction errors \_pb,m(t) and \_qb,m(t) becomes
\begin{align*}
E_{pb,m} &= r_{ys}(\theta) - q^T_{b,m}P_m^*P_{b,m}
\end{align*}

Due to the symmetries between \_f,m and \_b,m, \_p,m and \_q,m, the following symmetries can be established between the joint linear predictions: (1) \_f,m = \_b,p,m, (2) \_p,m = \_f,m, and (3) \_b,m = \_q,m. These symmetries are critical in reducing the complexity of the lattice-ladder structure. In addition, \_f,m and \_b,m in the subsequent discussion.

3.2. Order Adaptation
To construct a lattice-ladder structure, the prediction errors need to be computed recursively from order 1 to order M. Therefore, it is necessary to find the OAEs for the prediction errors which facilitate the computation of the prediction errors of order m+1 from those of order m.

Based on the definition of the prediction errors in Eqs. (6) and (9) and the ideal prediction vectors given in Eqs. (7) and (10), the OAEs for the prediction errors can be derived as
\begin{align*}
e_{pf,m+1}(t) &= e_{pf,m}(t) - H_m e_{pb,m}(t-1) \\
e_{pb,m+1}(t) &= e_{pb,m}(t-1) - K_m e_{pf,m}(t) \\
e_{qf,m+1}(t) &= e_{qf,m}(t) - K_m e_{qb,m}(t-1) \\
e_{qb,m+1}(t) &= e_{qb,m}(t-1) - H_m e_{qf,m}(t)
\end{align*}

where \_H,m and \_K,m are the lattice-ladder filter coefficients in the m-th stage with their ideal values being
\begin{align*}
H_m &= \left[ \begin{array}{c} -b_{pf,m}^T \v 1 \\
-1 ~ P_{f,m+1} \end{array} \right] \\
K_m &= \left[ \begin{array}{c} -b_{qf,m}^T \v 1 \\
-1 ~ P_{b,m+1} \end{array} \right]
\end{align*}
respectively. The identities in Eqs. (16) and (17) come from the symmetries (b_{pf,m} = \_f,m, b_{qf,m+1} = \_q,m, and (b_{pb,m} = \_f,m, \_b,p,m = \_q,m,\_b,q,m = \_p,m)).

To compute the output signal z(t) in the similar stage-by-stage style, the m-th order decorrelation filter is denoted as \_h,m with its solution being \_h,m = \_f,m (from Eq. (5)) and the stage output can be correspondingly defined as
\begin{align*}
z_{m}(t) &= d(t) - y_{f,m}^T P_{f,m}
\end{align*}
Therefore, the OAE for the stage output can be derived as
\begin{align*}
z_{m+1}(t) &= z_{m}(t) - L_m e_{pb,m}(t)
\end{align*}
where \_L,m is the third lattice coefficient in the m-th stage with its ideal value being
\begin{align*}
L_m &= \left[ \begin{array}{c} -b_{pf,m}^T \\
-1 \end{array} \right]
\end{align*}
The derivations of the OAEs are provided in [10]. Based on the definitions in Eqs. (6), (9), and (18), the input signals at the initial stage are e_{pf,0}(t) = e_{pf,0}(t) = y(t), e_{qf,0}(t) = e_{qf,0}(t) = v(t), and z_{0}(t) = d(t). Furthermore, by comparing Eqs. (4) and (18), the output signal at the final stage, z_{M}(t), is the desired output z(t). Eqs. (12)-(19) yield the lattice-ladder structure for the decorrelation filter described in Section 2.2. Fig. 1 shows the diagram of the derived lattice-ladder structure.

3.3. Estimation of the Lattice-Ladder Coefficients
In the derived lattice-ladder structure, \_H,m, \_K,m, and \_L,m in each stage need to be estimated. Therefore, the adaptive estimation equations (AEEs) for them need to be derived. If the lattice coefficients \_H,l, \_K,l, and \_L,l, l = 0, \ldots, m - 1 are all equal to their ideal values, it can be shown that
\begin{align*}
E_{pf,m} &= \left[ \begin{array}{c} -b_{pf,m}^T \v 1 \\
-1 \end{array} \right] P_{f,m+1} = E \{e_{pf,m}(t)e_{pb,m}(t-1)\}
\end{align*}
The ideal TIR is 5.51 dB in phones and the talkers, measured in a real acoustic environment on the direct-form and lattice-ladder ADF algorithms. The source speech signals are chosen from acoustic environment on the direct-form and lattice-ladder ADF algorithms. The source speech signals are chosen from the cross-interference between speech sources effectively. Two experiments are presented in this section. The first experiment compares the effects of changes in acoustic environment on the direct-form and lattice-ladder ADF algorithms. The source speech signals are chosen from TIMIT database. They are convolved with the impulse responses of acoustic paths and then added to generate the co-channel mixtures.

5.1. Performance of Lattice-Ladder ADF

In this experiment, the acoustic paths among the microphones and the talkers, measured in a real acoustic environment in the form of 200-sample (18.75 msc) impulse responses, were used to generate co-channel mixtures from source signals. The resulting target-to-interference ratio (TIR) is 5.51 dB in $y_1$ and 6.10 dB in $y_2$. After processed by the lattice-ladder ADF with 300 stages, the TIR was significantly improved to 14.84 dB in $v_1$ and 15.98 dB in $v_2$. An example of waveforms of $x_1$, $y_1$, and $v_1$ is shown in Fig. 3.

5.2. Changes in Acoustic Environment

In this experiment, the co-channel mixtures were generated from clean speech signals by using simulated acoustic paths. The ideal $f_{12}$ for direct-form ADF is shown in Fig. 4 and $\mathcal{H}_{m}$, $\mathcal{K}_{m}$, and $\mathcal{L}_{m}$ are shown in Fig. 5, all in solid
In the absence of adaptation, the TIR in after processing was 7.13 dB. To separate the co-channel mixtures generated after change, the filter coefficients were distributed more evenly in lattice-ladder ADF. It can be observed that the changes in the ideal filter coefficients in Figs. 4 and 5, respectively. From Figs. 4 and 5, it can be seen that the proposed algorithm is effective in reducing cross-interference between co-channel speech sources and can track the variations in acoustic environment better than direct-form ADF.

When the ideal filter coefficients before change were used to separate the co-channel mixtures generated after change without adaptation, the TIR in after processing was 7.13 dB in y1 and 24.06 dB in y2. When the impulse response of H12 was time-shifted to the left by 5 samples, although the TIR remained about the same (9.98 dB in y1 and 9.90 dB in y2), the ideal 1;1 and L1;1 changed significantly, as shown by the dashed curves in Figs. 4 and 5, respectively. From Figs. 4 and 5, it can be observed that the changes in the ideal filter coefficients distributed more evenly in lattice-ladder ADF.

When the ideal filter coefficients before change were used to separate the co-channel mixtures generated after change without adaptation, the TIR in after processing was 7.13 dB in y1 and 24.06 dB in y2 for both direct-form and lattice-ladder ADF. Here the negative impact caused by time-shifting in impulse response was obvious. When adaptation gains for both algorithms were chosen so that the TIR in v1 after processing are improved to around 22 dB after convergence, it can be observed from the learning curves in Fig. 6 that lattice-ladder ADF adjusted to the change faster than direct-form ADF.

6. CONCLUSION

This paper proposes a lattice-ladder ADF algorithm with the aim of developing a more efficient co-channel speech separation system. It is shown that based on the joint forward and backward linear predictions, a lattice-ladder structure can be developed for ADF. Experimental results show that the proposed algorithm is effective in reducing cross-interference between co-channel speech sources and can track the variations in acoustic environment better than the original direct-form ADF algorithm.

REFERENCES


