ROBUST AUTOMATIC SPEECH RECOGNITION USING A MULTI-CHANNEL SIGNAL SEPARATION FRONT-END

Kuan-Chieh Yen and Yunxin Zhao

Beckman Institute and Department of Electrical and Computer Engineering
University of Illinois at Urbara-Champaign

ABSTRACT

A multi-channel signal separation front-end for robust automatic speech recognition under time-varying interference conditions is developed. The speech signals acquired by a dual-channel system are restored by adaptive decorrelation filtering, and then examined by a time-domain or frequency-domain source signal detection technique to determine the active regions of each source signal. The front-end is integrated with an HMM-based speaker-independent continuous speech recognition system by providing the restored signals within the active regions for recognition. Under a simulated room acoustic condition, the overall system shows very promising performance. For the conditions with SNR above -10 dB, the achieved word recognition accuracies are very close to that of the interference-free condition.

1. INTRODUCTION

The state-of-the-art automatic speech recognition (ASR) techniques are still vulnerable in the presence of interference signal sources. The majority of the current research efforts on robust speech recognition has been focused on the reduction of stationary noises where the noise statistics either are known a priori or can be estimated from a certain inactive period of speech. [1,2,3] When the interference signals are time-varying, such as interference speech from a competing talker, the noise characteristics estimated at one instance of time are in general not applicable at a later time. In the current work, we propose using the adaptive decorrelation filtering (ADF) technique [4,5] as a front-end processing module for robust ASR under time-varying interference conditions.

Two microphones are used in the ADF algorithm; each microphones focuses on one signal source and, at the same time, acquires the interference signal from the other source. The signals processed by ADF have shown great improvement in our informal hearing evaluation. In integrating the ADF as a front-end processing module for ASR, however, it was found that the leakage signals (i.e. the interference signals remaining after ADF) still degraded recognition accuracy significantly in regions where the signal-to-interference ratio (SIR) was very low. A typical such case is that the speech signal of interest is absent and only the interference is present. To handle this problem, we take the approach of detecting the active regions of each source signals and performing speech recognition only on the interested speech signals within their active regions. We propose two methods to detect the presence of the source signals, where the methods are based on the second-order statistics of the ADF processed signals, one in time domain and the other in frequency domain.

We start this paper with a description of the dual-channel system and the decorrelation signal separation algorithm. We then introduce the methods of detecting the active regions for the source signals, and briefly describe the HMM-based speaker-independent continuous speech recognition system. Finally, we present experimental results and show the promising performance of the entire system.

2. THE DUAL-CHANNEL SYSTEM

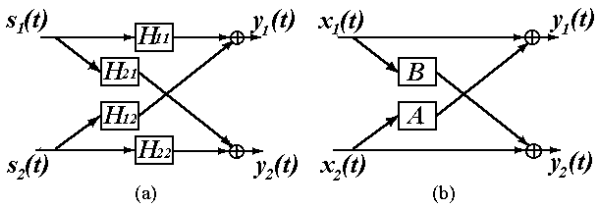

Figure 1: (a) The block diagram of the dual-channel system (b) The block diagram of the simplified dual-channel system

As illustrated in Figure 1(a), the observed signals ($y_1(t)$ and $y_2(t)$) acquired by the two microphones and the source signals ($s_1(t)$ and $s_2(t)$) in the dual-channel system are related by

$$Y_1(f) = H_{11}(f)S_1(f) + H_{12}(f)S_2(f)$$
$$Y_2(f) = H_{21}(f)S_1(f) + H_{22}(f)S_2(f)$$

where $H_{ij}(f)$ is the transfer function from the source i to the microphone j. We can rewrite the model as the following:

$$Y_1(f) = X_1(f) + A(f)X_2(f)$$
$$Y_2(f) = X_2(f) + B(f)X_1(f)$$

where

$$X_i(f) = H_{ii}(f)S_i(f), \quad i = 1, 2$$
$$A(f) = H_{12}(f)/H_{22}(f)$$
$$B(f) = H_{21}(f)/H_{11}(f)$$

i.e., $X_i(f)$ is a linearly distorted version of $S_i(f)$. Since our ASR system can handle linear channel distortion through acoustic normalization, we only focus on the recovery of the signals $x_i(t)$'s in the front-end processing. The simplified system is illustrated by Figure 1(b).

3. ADAPTIVE DECORRELATION FILTERING

In the system shown in Figure 1(b), the source signals can be perfectly restored if the filters A and B are known and $[1-AB]$ is nonzero for all frequencies. When A and B are unknown, the source signals can be estimated by the system shown in Figure 2, where A_e and B_e are the estimates of A and B, respectively, $\hat{x}_i(t)$ is the estimate of $x_i(t)$, $i = 1, 2$, and $C = [1 - A_e B_e]^{-1}$.

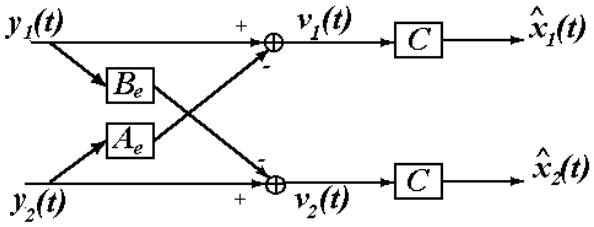

Figure 2: The block diagram of the signal separation system

The speech signals from different sources are assumed as zero-mean and independent, and hence the correlation between the two source signals is zero. Based on this assumption, assuming that A_e and B_e are FIR filters of orders N_a and N_b, respectively, the following adaptive algorithm can be used to estimate the filter coefficients of A_e and B_e: [4]

$$\mathbf{a}^{(t)} = \mathbf{a}^{(t-1)} + \gamma(t)\mathbf{v}_2^{(t-1)}(t)v_1^{(t-1)}(t)$$
$$\mathbf{b}^{(t)} = \mathbf{b}^{(t-1)} + \gamma(t)\mathbf{v}_1^{(t-1)}(t)v_2^{(t-1)}(t)$$

where

$$\mathbf{a}^{(t)} = [a^{(t)}(0) \ldots a^{(t)}(N_a)]^T$$
$$\mathbf{b}^{(t)} = [b^{(t)}(0) \ldots b^{(t)}(N_b)]^T$$
$$\mathbf{v}_1^{(t-1)}(t) = [v_1^{(t-1)}(t) \ldots v_1^{(t-1)}(t-N_b)]^T$$
$$\mathbf{v}_2^{(t-1)}(t) = [v_2^{(t-1)}(t) \ldots v_2^{(t-1)}(t-N_a)]^T$$

where $a^{(t)}(k)$ and $b^{(t)}(k)$ are the filter coefficients estimated at time t, $v_i^{(t-1)}(.)$ is the signal $v_i(.)$, $i = 1, 2$, based on the estimated filters at time $t-1$, and T denotes vector transpose.

In our experiments, the adaptation rate $\gamma(t)$ was chosen according to the amplitude of the observed signals in both channels, the orders of the FIR filters A_e and B_e, and the time-varying rate of the two channels.

4. SOURCE SIGNAL DETECTION

Although the ADF algorithm discussed above is efficient in canceling out the cross-talk in both channels, the leakage signals still pose problems in the automatic recognition of the restored signals. In regions where the interested speech source is inactive for an extended period of time, even a very weak leakage signal from the other channel can deteriorate the recognition accuracy. In the current work, our strategy is to detect the active and inactive regions of the source signals and perform speech recognition only within the active regions of each signal.

From the systems shown in Figures 1(b) and 2, the relationship between the restored signals and the source signals is:

$$\hat{X}_1 = \frac{1-A_eB}{1-A_eB_e}X_1 + \frac{A-A_e}{1-A_eB_e}X_2 \quad (1)$$
$$\hat{X}_2 = \frac{1-AB_e}{1-A_eB_e}X_2 + \frac{B-B_e}{1-A_eB_e}X_1 \quad (2)$$

where the first term in the right hand side of each equation is a distorted source signal, and the second term is a leakage signal. Assuming that the estimates of the channels A and B are very close to the true filters (i.e. $A_e \approx A$ and $B_e \approx B$), then the distortion is ignorable and the equations (1) and (2) can be simplified as:

$$\hat{X}_1 \approx X_1 + GX_2, \quad \text{with } |G| << 1, \ \forall f \quad (3)$$
$$\hat{X}_2 \approx X_2 + HX_1, \quad \text{with } |H| << 1, \ \forall f \quad (4)$$

i.e., each restored signal is approximately equal to its source signal plus a small leakage from the other channel. Based on the second-order statistics in either time domain or frequency domain, the following three hypotheses are made for each frame of the restored signals:

H_0: *Both speech sources are active.*
H_1: *Only the speech source in channel 1 is active.*
H_2: *Only the speech source in channel 2 is active.*

4.1. Time Domain Source Detection

By rewriting the equations (3) and (4) in time domain and use FIR filters to approximate G and H, we have

$$\hat{x}_1(t) = x_1(t) + \sum_{k=0}^{N_g} g(k)x_2(t-k), \quad \sum_{k=0}^{N_g} |g(k)| << 1$$
$$\hat{x}_2(t) = x_2(t) + \sum_{k=0}^{N_h} h(k)x_1(t-k), \quad \sum_{k=0}^{N_h} |h(k)| << 1$$

Assuming that the signals are stationary within periods of the lengths of the filters, then

$$var\{\hat{x}_1\} = r_1(0) + \sum_k \sum_l g(k)g(l)r_2(k-l)$$
$$var\{\hat{x}_2\} = r_2(0) + \sum_k \sum_l h(k)h(l)r_1(k-l)$$
$$cov\{\hat{x}_1, \hat{x}_2\} = \sum_k h(k)r_1(k) + \sum_k g(k)r_2(k)$$

where the autocorrelation functions of the source signals $r_i(k)$'s are defined as $r_i(k) = E\{x_i(t)x_i(t-k)\}$, $i = 1, 2$.

Based on the above analysis, we observe that if the energy levels of both source signals are comparable, i.e. $r_1(0) \sim r_2(0)$, then $cov\{\hat{x}_1, \hat{x}_2\}$ should be smaller than $var\{\hat{x}_i\}$, $i = 1, 2$, due to the attenuation introduced by the two filters $g(k)$'s and $h(k)$'s. On the other hand, if the energy of the source i is much stronger than that of the source j, i.e. $r_i(0) >> r_j(0)$, then $cov\{\hat{x}_1, \hat{x}_2\}$ should be smaller than $var\{\hat{x}_i\}$ and larger than $var\{\hat{x}_j\}$. Since a source is very likely to be inactive if it is much weaker than the other source for a long period of time, we can block each of the restored

signals into a sequence of frames and label each pair of frames of $\hat{x}_i$'s by D_1 and D_2 according to the following rule:

If $cov\{\hat{x}_1, \hat{x}_2\} < var\{\hat{x}_i\}$ for both $i = 1, 2$, set $D_1 = D_2 = 1$.
If $var\{\hat{x}_1\} > cov\{\hat{x}_1, \hat{x}_2\} > var\{\hat{x}_2\}$, set $D_1 = 1$ and $D_2 = 0$.
If $var\{\hat{x}_2\} > cov\{\hat{x}_1, \hat{x}_2\} > var\{\hat{x}_1\}$, set $D_1 = 0$ and $D_2 = 1$.

The D_i's are averaged over neighboring frames and then thresholded to 0 (inactive) or 1 (active), to label the active regions of each channel.

4.2. Frequency Domain Source Detection

Writing the equations (3) and (4) in terms of the short-time DFT of length N, we have

$$\hat{X}_1(k) = X_1(k) + G(k)X_2(k)$$
$$\hat{X}_2(k) = X_2(k) + H(k)X_1(k)$$

where $|G(k)| << 1$ and $|H(k)| << 1$ for $k = 0, ..., (N-1)$.
Define the correlation coefficient ρ_k for each frequency bin k as

$$\rho_k^2 = \frac{E\{\hat{X}_1(k)\hat{X}_2^*(k)\}E\{\hat{X}_1^*(k)\hat{X}_2(k)\}}{E\{|\hat{X}_1(k)|^2\}E\{|\hat{X}_2(k)|^2\}}$$

In each frequency bin, $X_1(k)$ and $X_2(k)$ are zero-mean and independent. Let $E_{i,k} = E\{|X_i(k)|^2\}$ and $E_i = \sum_{k=0}^{N-1} E_{i,k}$, $i = 1, 2$, then

$$E\{|\hat{X}_1(k)|^2\} = E_{1,k} + |G(k)|^2 E_{2,k}$$
$$E\{|\hat{X}_2(k)|^2\} = E_{2,k} + |H(k)|^2 E_{1,k}$$
$$E\{\hat{X}_1(k)\hat{X}_2^*(k)\} = H^*(k)E_{1,k} + G(k)E_{2,k}$$
$$E\{\hat{X}_1^*(k)X_2(k)\} = H(k)E_{1,k} + G^*(k)E_{2,k}$$

In the k-th bin, if $E_{1,k} \approx E_{2,k}$, then $\rho_k^2 \approx 0$. On the other hand, if $E_{1,k} >> E_{2,k}$ or $E_{2,k} >> E_{1,k}$, then $\rho_k^2 \approx 1$. If the energy of a source signal is much weaker than that of the other in most frequency bins for an extended period of time, then this source is very likely to be inactive. Therefore, by defining $P = \sum_{k=0}^{N-1} \rho_k^2$ and choosing a threshold value T, the signal frames can be labeled according to the following rule:

H_0: if $P \le T$.
H_1: if $P > T$ and $E_1 > E_2$.
H_2: if $P > T$ and $E_2 > E_1$

The expectations $E\{\hat{X}_i(k)\hat{X}_j^*(k)\}$ are approximated by blocking each signal into a sequence of frames and averaging the DFT coefficients $\hat{X}_i(k)\hat{X}_j^*(k)$ over $2M+1$ neighboring frames.

5. AUTOMATIC SPEECH RECOGNITION

The speaker-independent continuous speech recognition system is based on the hidden Markov models of phone units: each phone-unit HMM has three tied-states; each state is modeled by a Gaussian mixture density. For each mixture density, the basis Gaussian densities are context-independent; the mixture weights are triphone context-dependent; the mixture size and the Gaussian density parameters are determined via a bottom-up merging algorithm. The phone models were trained from a subset of 717 sentences in the TIMIT database, with a total of 62 units as defined in the TIMIT (excluding "h#") [6]. The average mixture size per mixture density is approximately 19, and the total number of triphone context is about 4500. The TIMIT speech data were down-sampled from 16 KHz to 10.67 KHz. The cepstrum coefficients of the PLP analysis (8th order) and log energy were taken as instantaneous features and their first-order 50-msec temporal regression coefficients as dynamic features. Further details of the background materials can be found in [7]. The acoustic normalization technique was based on cepstral bias estimation which was bootstrapped by cepstral mean estimation. Details can be found in [8].

6. EXPERIMENTS

A subset of TIMIT database which forms 78 sentence pairs was chosen as the test set for source signals. The signals were scaled to obtain relative source energy levels (RSEL) between two channels: 0 dB, ±10 dB, and ±20 dB. The cross-talk coupling filters were FIR filters which simulated two microphones located one meter apart and 10 cm away from their respective sound sources. These filters introduced attenuation of about 8.37 dB. The speech recognition task has a vocabulary size of 835 and grammar perplexity of 105.

6.1. Adaptive Decorrelation Filtering

In our experiment, the ADF algorithm was found to be capable of improving the source-to-leakage ratio (SLR) in both channels for all RSEL cases. It was also judged favorably in the listening tests. The improvement was very impressive in the channel with low RSELs. An example is in Figure 3(a-c) which illustrates the effectiveness of this algorithm.

6.2. Time-Domain Source Detection

For source detection, the restored signals were divided into frames of 800 samples with steps of 160 samples. The values were averaged over the neighboring 21 frames. The detection error rates are summarized in Table 1, where labeling a source as active when the source was inactive is defined as the false alarm rate, and labeling a source as inactive when the source was inactive is defined as the miss rate.

RSEL	20 dB	10 dB	0 dB	-10 dB	-20 dB
False alarm rate (%)	30.04	17.06	11.16	11.13	9.17
Miss rate (%)	0.03	0.15	0.28	0.74	3.82

Table 1: The detection error rates of the time-domain source detection technique

6.3. Frequency-Domain Source Detection

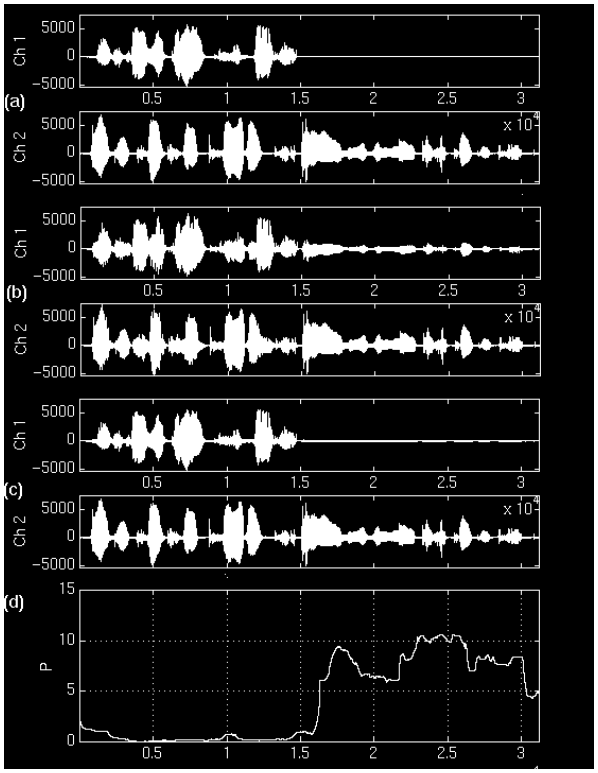

Figure 3: An example: (a) The source signals [SOUND_A425S01.WAV] [SOUND_A425S02.WAV] (b) The observed signals [SOUND_A425S03.WAV] [SOUND_A425S04.WAV] (c) The restored signals [SOUND_A425S05.WAV] [SOUND_A425S06.WAV] (d) The plot of *P* in frequency-domain source detection

The plot of *P* for the signal pair in Figure 3(a-c), using a length-64 DFT and a 3201-sample window in approximating the expectation values, is shown in Figure 3(d). From the plot we see a sharp rise of *P* at about 1600 samples after the source signal in channel 1 stopped, due to the window size we used.

6.4. Automatic Speech Recognition

The recognition word accuracies of the following types of signals are evaluated and summarized in Table 2:

1. Test set source signals
2. Observed signals
3. Restored signals generated by the ADF algorithm
4. Restored signals using hand-labeled endpoints
5. Restored signals processed by time domain detection
6. Restored signals processed by frequency domain detection

RSEL	20 dB	10 dB	0 dB	-10 dB	-20 dB
Signals in type 1 (%)	91.20	91.20	91.20	91.20	91.20
Signals in type 2 (%)	68.90	59.10	20.0	-20.50	-20.60
Signals in type 3 (%)	72.30	72.50	73.60	71.80	59.40
Signals in type 4 (%)	91.60	90.80	88.10	85.60	73.30
Signals in type 5 (%)	85.90	89.60	87.60	86.40	70.30
Signals in type 6 (%)	91.10	90.0	86.10	84.20	68.40

Table 2: Summary of the recognition word accuracy

From Table 2, the recognition accuracies of the signals after source signal detection are close to the accuracy of source signals in most cases except the case of -20 dB RSEL. In this case the accuracy is also close to the one with hand-labeled endpoints. Overall, the integration of the ADF and source signal detection improved the recognition performance significantly.

7. CONCLUSION

Our current work demonstrates that the signal separation front-end is very promising for robust ASR under time-varying interference conditions. The recognition accuracies of processed signals are comparable to that of the source signals except for those with very low SIR in the observed signals. Further improvement in recognition accuracy requires better restoration filtering which is currently under investigation. The evaluation of our system performance under real acoustic conditions is also underway.

8. ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. IRI-95-02074.

9. REFERENCES

1. B.-H. Juang, "Speech Recognition in Adverse Environments," *Computer Speech and Language*, pp. 275-294, May, 1991.
2. A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech Recognition Using Noise-Adaptive Prototypes," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 37, pp. 1495-1503, Oct. 1989.
3. M. J. F. Gales and S. J. Young, "An Improved Approach to the Hidden Markov Model Decomposition," *Proc. ICASSP*, pp. 729-734, 1992.
4. E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-Channel Signal Separation by Decorrelation," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 4, pp. 405-413, Oct. 1993.
5. S. Van Gerven and D. Van Compernolle, "Signal Separation by Symmetric Adaptive Decorrelation: Stability, Convergence, and Uniqueness," *IEEE Trans. on Signal Processing*, Vol. 43, No. 7, pp. 1602-1612, Jul. 1995.
6. L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. of Speech Recognition Workshop (DARPA)*, 1986.
7. Y. Zhao, "A Speaker-Independent Continuous Speech Recognition system Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 3, pp. 345-361, Jul. 1993.
8. Y. Zhao, "Self-Learning Speaker and Channel Adaptation Based on Spectral Variation Source Decomposition," *Speech Communication*, Vol. 18, pp. 65-77, Jan., 1996.