DOUBLETALK DETECTION USING REAL TIME RECURRENT LEARNING

[1]*Mohammad Asif Iqbal,* [2]*Jack W. Stokes,* [2]*John C. Platt,* [2]*Arun C. Surendran and* [1]*Steven L. Grant*

[1]ammq2@umr.edu
[1]University of Missouri-Rolla, Rolla, MO 65409
[2]Microsoft Research, Redmond, WA 98052

ABSTRACT

In this paper we present a new system for doubletalk detection that uses multiple signal detectors/discriminators based on recurrent networks. The goal is to build a simple system that learns to combine information from different signal sources to make robust decisions even under changing noise conditions. In this paper we use three detectors - two of these are frequency domain signal detectors, one at the far-end and one at the microphone channel. The third detector determines the relative level of near-end speech vs far-end echo in the microphone signal. The new double-talk detector combines information from all these detectors to make its decision. An important part of this proposed design is that the features used by these detectors can be easily tracked online in the presence of noise. We compare our results with cross-correlation based doubletalk detectors to show its effectiveness.

1. INTRODUCTION

Acoustic echo cancelers (AEC) are important part of teleconferencing systems - they are necessary to mitigate the deleterious effect of acoustic feedback from the speaker signal to the microphone input [1]. In an AEC, the echo path is adaptively modeled using a filter, which is then used to synthesize a replica of the echo and subtract it from the echo-corrupted microphone signal [2]. When the near-end talker is active, or when there is no far-end signal, the filter coefficients will diverge from the true echo path impulse response; hence is it crucial to have a good *doubletalk detector* which indicates periods of simultaneous far-end and near-end speech. During these periods the adaptation of the filter coefficients is stopped [1].

Double-talk detection can use statistics computed from the both the microphone- and the far-end signal. Typically a cross-correlation based statistic is used in these scenarios [3]. In addition, some statistics based on each individual signal may also be computed which can assist in the detection. It may not be straighforward to analytically compute and combine all this information based on correlation analysis alone; in this paper we propose a machine learning based approach.

In our new approach, we propose to use multiple speech detectors/discriminators (D/D) at various points, and then combine them for effective doubletalk detection. The system is modular in nature, so it is extendable to multi-channel scenarios. But in this paper we demostrate the idea on a system with a single microphone channel. In this system, we use three different D/D units. Two of them are signal detectors and are used to detect the presence of a signal at the far-end (FESD) and at the near-end (NESD) as shown in Figure 1. At the near-end, the signal can be due to near-end speech or due to echo from the far-end talker. Thus we need a third unit, which is a discriminator - it estimates the relative influence of far-end echo vs the near-end speech in the microphone signal. For lack of a better term, we call this third unit simply "signal discriminator" (SD). The final part of our double-talk detector combines the output of all these units to make robust decision regarding double-talk.

Since the detectors have to be robust to changining noise conditions, we propose to use SNR dependent features which have been shown to be effective for speech detection [4], and can be easily tracked online in the presence of noise.

This paper is structured as follows: In section 2 we present our method for singal detectors/discriminators and for doubletalk detection. In section 3 we discuss the experiments and results which is followed by a summary and conclusion in section 4.

2. SIGNAL DETECTORS/DISCRIMINATORS

One of our primary goals is to make the overall system have low complexity - this requires that the D/D units themselves be very simple. Recently logistic [4] networks were shown be very simple and effective for speech detection even in changing noise conditions. This idea can be easily carried over to detecting other types of signals in noise.

In our acoustic application, all the signals are influenced by reverberation, whose effect typically lasts for hundreds of milliseconds; further speech itself is a highly correlated signal. Hence it is important that our detectors incorporate this long-term effect in them automatically. One way to achieve this is to take multiple frames of data (spanning the desired time-length of interest) and use them as inputs to the network. One problem with this approach is that the correct number to include will depend upon the situation, and will have to be determined by trial and error. This also makes the network more complex. Another option is to use past decisions rather than features. *Recurrent networks* [2] are excellent examples of systems that achieve this - they dynamically re-use information about the state of the network from the past (these typically constitute the previous outputs of the network) as inputs to the current decision.

Combining the above two ideas, we propose to use a single layer network with recurrent feedback (shown in Figure 3). The state space model of our system can be written as:

$$x(n) = (1-\alpha)(\sum_{i=1}^{N} w_i u_i) + \alpha x(n-1) \quad (1)$$

$$y(n) = \frac{1}{1+\exp(-x(n))} \quad (2)$$

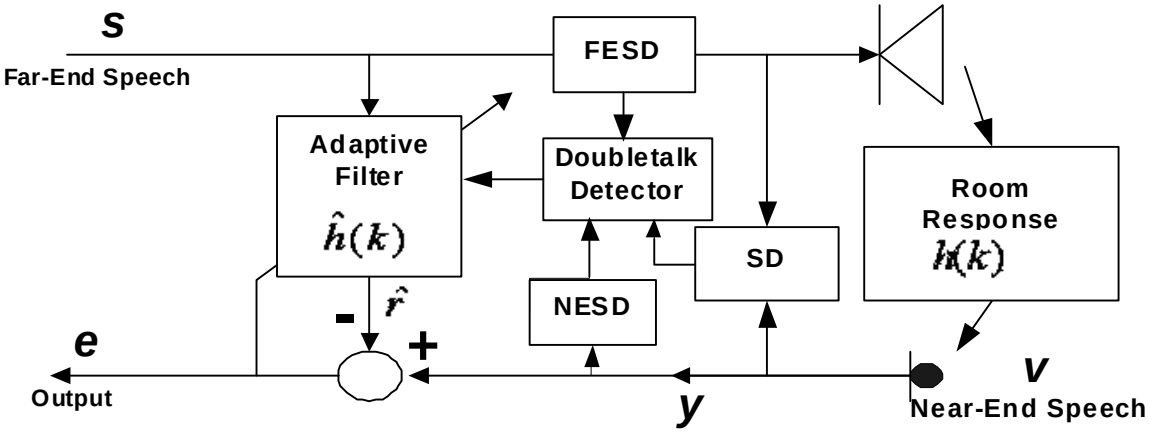

Figure 1: An AEC system showing various modules of our double-talk detector

where $[u_1(n)u_2(n)\ldots u_{N-1}(n)1]$ is the current input data and w_is and α are the parameters of the system. $y(n)$ is a value between 0 and 1, and hence can be interpreted as a probability. Since the input features are time-dependent, and arrive one per time-segment, it is appropriate to train this network continuously in on-line fashion after every frame of data arrives. This type of learning is appropriate for a non-stationary signal like speech, and is called *real-time recurrent learning* (RTRL) [5]. RTRL uses stochastic gradient descent to train this network to minimize the cross-entropy error [6]. This error metric makes the network discriminative, and provides the maximum likelihood estimate of the class probability for a wide variety of class conditional densities of the data [6]. The reason this is useful for us is that, since the outputs represent probabilities, it is easy for us to make decisions based on them, or combine their decisions with others. Further details of the training can be found at [**?**].

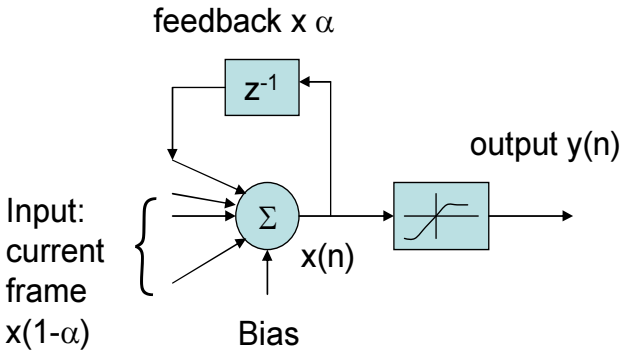

Figure 2: Recurrent network architecture

2.1. Feature design

One of the desired characteristics of any detector is that its features are sufficiently simple, easily to calculate, have discriminatory power and work well under changing noise conditions. We use estimated posterior SNR $\chi(k,t)$ as the feature set for the NESD and FESD (these have been shown to have all the above desirable properties [4]). $\chi(k,t)$ is the ratio of the energy in a given time-frequency atom S to the noise energy N $\chi(k,t) = \frac{|S(k,t)|^2}{N(k,t)}$ where k,t are the frequency bin and time indices respectively. the FESD uses the speaker signal S as the target signal, and the NESD uses the microphone signal Y. The short term spectra of speech are modeled well by log-normal distributions; hence we use the logarithm of the SNR estimate rather than the SNR estimate itself. Thus the inputs used are:

$$\chi_{FESD}(k,t) = \{\log|S(k,t)|^2 - \log N_{FE}(k,t)\} \quad (3)$$

and

$$\chi_{NESD}(k,t) = \{\log|Y(k,t)|^2 - \log N_{NE}(k,t)\} \quad (4)$$

where N_{FE} and N_{NE} are the noise energies in frequency bin k and time-frame t at the far-end and near-end respectively. The noise power N can be tracked using various algorithms such as [7, 8]. In this paper we use a minima tracker (for each frequency bin we look back a few frames e.g. 25, and choose the lowest value of the signal) followed by smoothing, to track the noise floor [8].

We describe the features for the speech discriminator (SD) next. SD is trying to look at the microphone signal, and it is trying to figure out how much of it is dominated by the near-end speech (as opposed to the far-end echo). Thus it is trying to discriminate the level of near-end speech. Thus for this system we use the logarithm of the ratio of the microphone instantaneous power Y to the far-end instantaneous power S for each frequency bin per frame as the feature i.e.

$$\chi_{SD}(k,t) = \log|Y(k,t)|^2 - \log|S(k,t)|^2. \quad (5)$$

As can be seen in Figure 2, the extracted features are clearly distinct for different scenarios. As expected, the extracted features are typically largest for only the near-end speech, smallest for the echo-only case, and in between for the case of doubletalk. Different feature levels correspond to different probability levels; larger features correspond to higher probabilities. For the echo-only case, the extracted features are always low independent of the echo-path; hence the discriminator performance is relatively independent of the echo-path. We have verified this empirically

under a wide variety of situations. The decision from this discriminator by itself is not very accurate for double-talk detection, but we hope to make better decisions when combined with decisions from NESD and FESD. It is probably best to build another learner which combines all these three decisions into one. In this paper, we use a simple approach (outlined below). In future works we hope to improve upon this.

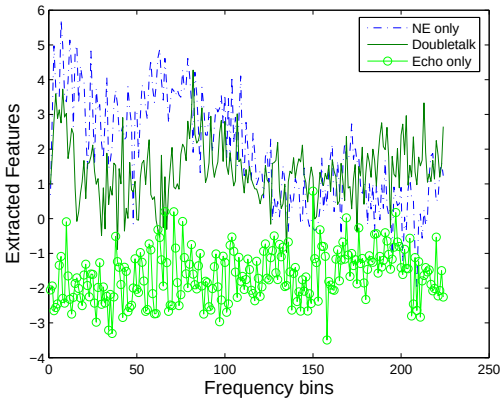

Figure 3: Extracted Features for the SD.

When the NESD and the SD of Figure 1 both indicate a high probability of the presence of speech, above the selected threshold, we confirm the presence of near-end speech. If the FESD of Figure 1 indicates the presence of speech and we have a confirmed near-end talker, then we declare the current-frame of the captured signal to be doubletalk. In short, we declare doubletalk when all the three detectors indicate the presence of speech; we declare the presence of near-end speech when NESD and SD detect speech while the FESD indicates a low probability of speech.

3. EXPERIMENTS AND RESULTS

We use the well known AURORA database [9] for our experiments. The recorded digital speech is sampled at 16 KHz and is used for the far-end speech *s* and the near-end speech *v* of Figure 1. We measured the room impulse response of a $10' \times 10' \times 8'$ room using a stereo system; the truncated 8000 sample (500 ms) room response is used as the loudspeaker-microphone environment *h* in Figure 1. A subset of the Aurora data base was used for training the FESD of Figure 1 precisely 75 signals (50000 frames) consisting of a mixture of male and female speakers. These signals were filtered through the left channel of the measured room impulse response to create the echo part of the microphone signals; near-end speech signals (different signals taken from the Aurora database) were added to simulate the microphone signals for training the NESD and the SD of Figure 1. Near-end speech was added at different near-end to far-end ratios to improve training.

For testing we use a completely different set of 120 signals taken from the Aurora data-base [9] to simulate the far-end speech. These signals were filtered using the right channel of the measured room impulse response to simulate a different channel for testing. To these artificially created echo signals we add near-end speech from a second different set of 120 signals taken from Aurora data-base at 12 different near-end to far-end ratios (NFR). We, thus have ten signals for testing at each NFR ratio where each signal is approximately 8-10 seconds long.

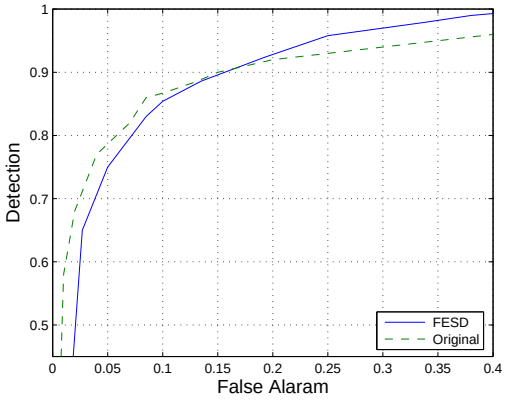

Figure 4: ROC Curve for the FESD, Original curve taken directly from [4].

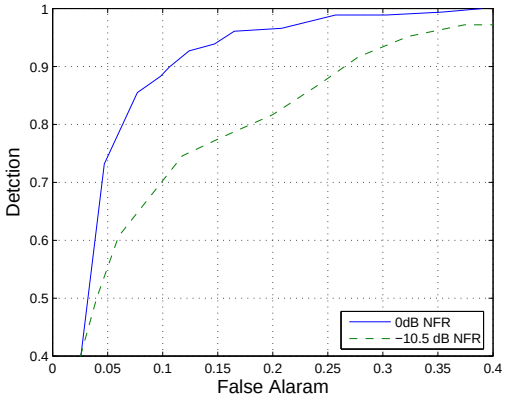

Figure 5: ROC Curve for Detecting NE Speech at Different NFR.

The true labels on the speech signals were generated by thresholding the energy in each time frame of the clean data; the threshold was selected so that all the speech events were retained, which was verified by listening to a small fraction of the training data. To study the performance of the speech detectors we plot the ROC curves (correct detection of speech versus false alarm). As can be observed from Figure 4, results are compatible with the speech detector of [4] which was trained with 8 KHz sampled speech . As a result, we confer that the training is done appropriately for the FESD.

The presence of near-end speech is confirmed when both the NESD and the SD indicate presence of speech. We combine both the NESD and the SD and plot the ROC curve in Figure 5 at different values of NFR. At a false alarm rate of 0.1, we detect the near-end speech with a detection probability of 0.89 at 0 dB NFR; as expected we detect the near-end speech with a lower detection rate of 0.7 at -10.5 dB NFR. We can clearly observe that

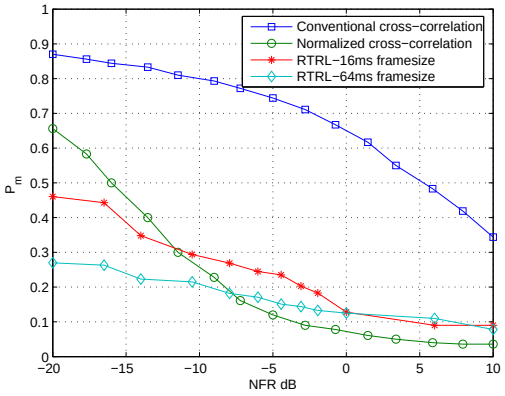

Figure 6: P_m as function of NFR for doubletalk detectors using our method, normalized cross-correlation based detector and the conventional cross-correlation based detector at $P_f = 0.1$.

we have a better detection rate at 0 dB as compared to -10.5 dB NFR as should be the case. The axes are truncated to highlight the upper left quadrant of the plot.
To obtain the thresholds corresponding to $P_f = 0.1$ (probability of false alarm = 0.1), we follow [3]:

1. Set $v = 0$ (No near-end speech).
2. Select thresholds for all the speech detectors.
3. Compute P_f.
4. Repeat steps 2, 3 over a range of threshold values.
5. Select the thresholds that correspond to $P_f = 0.1$.

These thresholds were used to compute the probability of miss, P_m , for the test signals. For the ten signals at each NFR, we average the P_m over the respective signals to calculate the average probability of miss P_m.
For evaluating the new RTRL doubletalk detector we closely follow [10]. Results are compared with the new normalized cross-correlation based detector [3] and the conventional cross-correlation based detector [11]. The P_m characteristics of all three methods under the constraint of $P_f = 0.1$ are shown in Figure 6. The RTRL doubletalk detector proposed here clearly outperforms the conventional cross-correlation based detector over a full range of NFR. Our new algorithm outperforms the normalized cross-correlation based detector for lower values of NFR and is comparable over the remaining region. It must be noted that we work with a frame size of 16 ms (256 samples at 16 KHz) whereas the other methods use a frame of size 62.5 ms (500 samples at 8 KHz).
Next we implement a bi-level architecture by aggregating 4 frames into a single frame so as to have a frame of duration 64 ms comparable to that of the normalized cross-correlation based detector's 62.5 ms. We observe in Figure 6, that the RTRL doubletalk detector outperforms the normalized cross-correlation based detector in almost half of the range of NFR values and is very close in the remaining region.
The FESD has a detection rate of 0.88 at 15 dB SNR (Figure 4); thus the RTRL based doubletalk detector is bounded by a miss probability of 0.1 even at higher NFR values (Figure 6). Typically in a teleconferencing device such as the Microsoft RingCam [12] the loudspeaker is located very close to the microphone, and the near-end talkers are relatively further away from the microphone. Thus, we typically have low NFR values in such devices. As can be observed from Figure 6, the RTRL based doubletalk detector significantly outperforms the normalized cross-correlation based detector over such lower NFR values making it suitable to use for such applications. Computational complexity of the RTRL doubletalk detector is of the order of L, whereas for the correlation based detectors it is of the order of $L \log L$,where $L = 256$ samples is the frame length.

4. CONCLUSION

We have proposed a new doubletalk detector based on a novel near-end speech detector; we significantly outperform the conventional cross-correlation based detector and are comparable to the normalized cross-correlation based detector.
Echo is a delayed speech signal; typically the spectrum of the echo is very similar to the spectrum of a speech signal with a quicker falloff from the maxima. Since we work in the frequency domain, we observe that the trained coefficients are equally applicable to any room responses. Similar results were observed for different room responses and even better results were observed with real data collected using the RingCam project at Microsoft Research [12]. Based on these observations we conclude that the trained weights are equally applicable to any room responses if not independent of room responses.

5. REFERENCES

[1] J.Benesty, T. Gansler, D.R. Morgan, M.M. Sondhi, and S.L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer, Inc., New York, 2001.

[2] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs, NJ, 1991.

[3] Jacob Benesty, Dennis R. Morgan, and Juan H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 168–172, March 2000.

[4] Arun C. Surendran, Somsak Sukittanon, and John Platt, "Logistic discriminative speech detectors using posterior snr," in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, pp. 625–628.

[5] Ronald J. Williams and David Zipser, "Experimental analysis of real-time recurrent learning algorithm," in *Connection Science,Vol 1, No 1*, 1989, pp. 87–111.

[6] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press.

[7] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," in *Signal Processing 81*, 2001, pp. 2403–2418.

[8] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of the 7th European Signal Processing Conference*, Edinburgh , Scotland, September 1994, pp. 1182–1185.

[9] David Pearce and H.G. Hirsch, "The aurora experimental framework," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, October 2000, pp. 16–20.

[10] Juan H. Cho, Dennis R. Morgan, and Jacob Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 718–724, November 1999.

[11] Hua Ye and Bo-Xiu Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Transactions on Communications*, vol. 39, pp. 1542–1545, November 1991.

[12] Ross Cutler, "The distributed meetings system," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, China, April 2003, pp. 756–759.