SUBBAND ECHO CANCELLATION IN AUTOMATIC SPEECH DIALOG SYSTEMS

Andrej Miksic and Bogomir Horvat
Laboratory for Digital Signal Processing
Faculty of Electrical Engineering and Computer Science
University of Maribor, Smetanova 17, 2000 Maribor, Slovenia
Tel. +386 62 221112, E-mail: andrej.miksic@uni-mb.si

ABSTRACT

Echo cancellation has been most widely studied for hands-free telephony and for cancelling line echos in telephone central offices. The problem of echo cancelling in speech dialog systems is similar, however it has some specific requirements.

In this contribution, a subband echo cancellation structure is proposed which can be integrated in the feature extraction part of a recognizer. A NLMS gradient-based adaptation is performed in frequency subbands that can either be derived directly from FFT analysis of input speech signal, or by using a proposed reduced-subband approach where the number of subbands is reduced in order to lessen the aliasing effect of the FFT.

A double-talk detector is proposed based on the estimated error function for decision on stopping the adaptation. Finally, a new approach of combining echo cancellation and noise reduction is proposed.

1. INTRODUCTION

In contrast to echo cancellation schemes used in man to man communication systems, where telephone line echoes are cancelled in telephone central offices and acoustic echoes are cancelled in hands-free telephony, the purpose of an echo canceller used in man-machine telephone dialog systems is to allow a better man-machine interaction by allowing a speaker (user of the dialog system) to utter to the telephone set without being forced to wait for the end of a system prompt uttered by the machine (talk-through capability).

Because of echoes generated in telephone hybrid circuits, a sampled received speech that is an input to a feature extraction part of a recognizer contains not only a speech signal from the user, but also an echo signal of a system prompt (double-talk situation), thus greatly degrading the recognition performance (figure 1).

Echo canceller generates a replica of the echo signal and subtracts it from the received speech. The residual error signal is commonly used to update the echo-canceller coefficients vector such that the mean squared value of the output error in minimized (LMS algorithm [4]).

In the feature extraction part of a speech recognition system, the speech spectra is commonly calculated. This gives a motivation to perform the echo cancellation on frequency subband signals.

2. A FREQUENCY SUBBAND ECHO CANCELLATION

Sub-band adaptive filtering is also one of the solutions to increased computation and slow convergence associated with the conventional full-band approach [1]. Reduced computational load is due to a time decimation in sub-bands (time/frequency transformation is performed in frames, not after each sample, so there is time decimation by down-sampling factor R). The subband echo canceller part of figure 2 (without band reduction and band recomposition) represents a block diagram of such a system.

With the reference to figure 2, the equations for NLMS algorithm performing the subband echo cancellation are the following:

$$\hat{Y}_i(l) = \sum_{m=0}^{M-1} c_{im}(l) X_i(l-m) \qquad \text{i=0..N-1} \qquad (1)$$

$$e_i(l) = Y_i(l) - \hat{Y}_i(l) \qquad \text{i=0..N-1} \qquad (2)$$

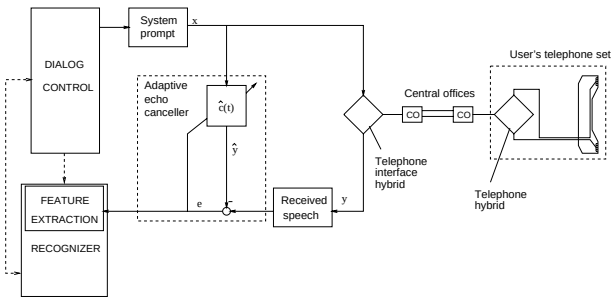

Figure 1: General structure of an echo canceller used in a telephone dialog system.

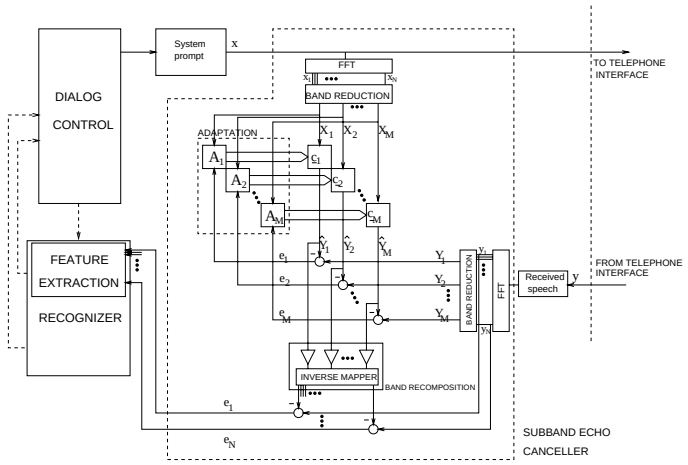

Figure 2: A reduced-subband echo cancellation. Frequency subbands are also inputs to a recognizer, therefore no re-synthesis of the received speech signal is needed.

$$c_{im}(l+1) = c_{im}(l) + 2\alpha \frac{e_i(l) X_i(l-m)}{\sum_{j=0}^{M-1} X_i^2(l-j)} \qquad \text{i=0..N-1} \tag{3}$$

where $\hat{Y}_i$ is the estimated echo at subband i (there are N subbands), M is the number of taps of the finite impulse response of the LTI system that estimates the telephone line (complete echo path), c_{im} are the coefficients of the LTI system that simulate the telephone line impulse response and e is the error in matching $\hat{Y}$ with Y. The system coefficients adapt and in the ideal case the error signal converges towards zero (convergence is possible when only echo is present – there is no user speech). Factor α is the step size factor of the NLMS algorithm.

3. A REDUCED-SUBBAND APPROACH

An echo canceller for man-machine dialog systems has specific requirements. It should be utilized in a way not to add a significant computational overhead to a speech recognizer and to avoid degradation in recognition performance. However, here is no need to resynthesize the received speech signal (see figure 1). The frequency subband signals can be led directly to the feature extraction part of a recognizer.

To reduce the computational load of the subband canceller and to assure a better stop-band attenuation (less aliasing needed for NLMS type adaptive filtering) of a subband division means, a band reduction is utilized in order to reduce the number of sub-bands from N (after the FFT of length N) to M (after the band reduction). This is possible because there is no need to reconstruct the original signal from the M subband signals. Only the feature set should be reconstructed (figure 2). Number of subbands M can be chosen even lower than the down-sampling factor R and the convergence of the NLMS algorithm is not obstructed.However, to preserve the spectral resolution of the received signal, the short-term magnitude spectrum of the received signal, obtained by FFT means must represent an input to a feature extraction such that the influences of echo energies from each of the N frequency components (i.e. magnitude spectrum components) must be subtracted. To achieve this, the band recomposition means are performed on the reduced number M of estimated subband echo energies to obtain the approximated echo influences on each of the N original subband signals.

Band reduction is performed by mapping a set of N frequency components $\{x_k, k = 1..N\}$ to M subband signals $\{X_i, i = 1..M\}$ where $M < N$. The mapping of indexes is defined by the mapping table $i = i(k)$ that defines which frequency components x_k are mapped into each subband signal X_i. The magnitudes of the subband signals X_i are calculated using this formula

$$X_i = \sqrt{\sum_k x_k^2},$$

where the sum is taken over all frequency components x_k that fall into band i according to the $i = i(k)$ mapping. Each subband i has a width w_i corresponding to the number of frequency components mapped into the subband X_i.

Band recomposition is performed by first passing the subband signals $\hat{Y}_i$ which represent the estimated echo magnitudes of band i through attenuation means that multiplies the subband signals of band i by the factor $\frac{1}{\sqrt{w_i}}$, where w_i is a width of the subband as described above. The inverse mapping now reconstructs the original N frequency channels according to the inverse of the index mapping table $i = i(k)$.

The magnitude spectrum components that reach the recognizer are therefore

$$e_k = y_k - \frac{\hat{Y}_{i(k)}}{\sqrt{w_{i(k)}}}$$

where $i(k)$ is the above mentioned index mapping table and $k = 1..N$.

With a function $i(k)$ we can map all possible divisions of 2^n frequency bands into subbands. In the results discussed later, we chose $i(k)$ mapping according to mel-scale. A reduction to 24 bands was achieved ($M = 24$).

The adaptation is the same as in equations 1, 2 and 3, only the number of subbands was reduced from N to M.

Comparing figures 5 and 6 (all figures are time-aligned), we can see how the error signals (input speech minus estimated echo) are less scattered in case of reduced-subband approach.

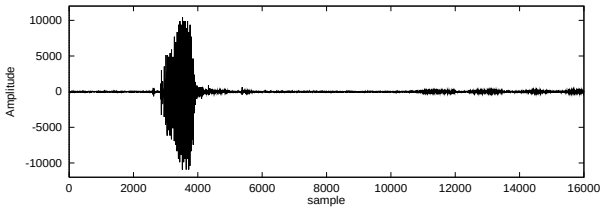

Figure 3: Utterance acht761.08 from the VM_5 database (user speech signal).

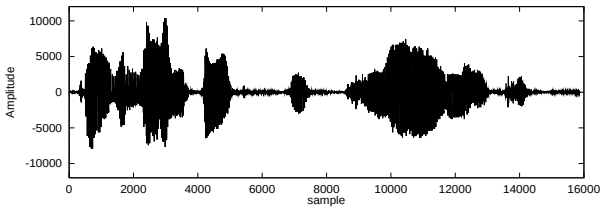

Figure 4: Echo signal "Die Mailbox des Teilnehmers ist voll...".

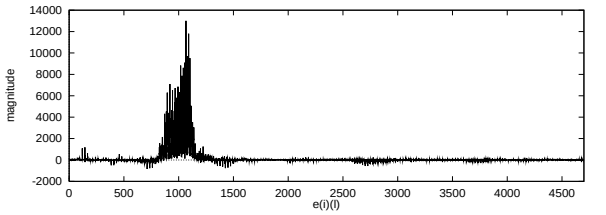

Figure 5: Error signals $e_i(l)$ for reduced-band approach. We can see how echo spectra is cancelled while user spectra enters the recognizer

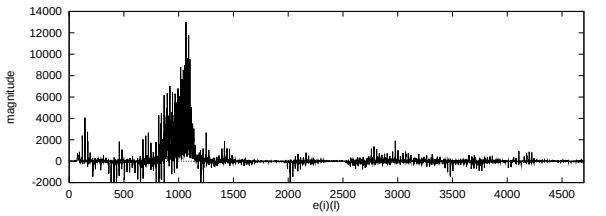

Figure 6: Error signals $e_i(l)$ for FFT band division approach. The residual spectra is too scattered becouse of aliasing.

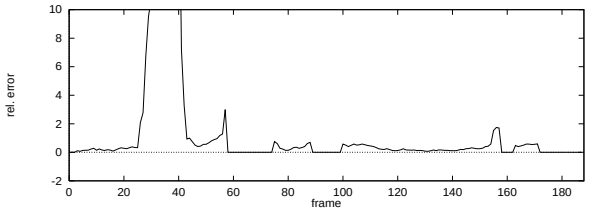

Figure 7: Relative error coefficient e_r used for double-talk decision.

4. STEPSIZE CONTROL - HANDLING THE DOUBLE-TALK CONDITIONS

It is important to define the system behaviour in the double-talk situations. On a dialog-level, two schemes are possible: when double-talk is detected, the system prompt could immediately be stopped and the system would only "listen" to the user. Such a system can lead to a very unpleasant dialog– if the double talk was wrongly detected when the user said nothing (or at least nothing useful) then the system question would stop and both parties would wait. We focused on a better and more complex solution: the system question stops only when the system has recognized a word or a sentence – only then it stops and asks for the confirmation.

The double-talk situation has to be detected in order to prevent missadjustment of the canceller. The adaptation should be frozen in case of a double-talk (stepsize α is set to zero). Methods for detecting the double-talk can be found in the literature (e.g. [7]) and the algorithms mostly use some kind of threshold values for the attenuation of the system question output power detected at the input of the automatic dialog system.

Based on experiments, we concluded that such a static double-talk detection method is not sufficient for the purposes of the automatic telephone dialog system. The most pronounced double-talk detection parameter was the error, relative to the current estimated input speech $\hat{Y}$:

$$e_r = \frac{\sum_{i=0}^{N}(Y_i - \hat{Y}_i)}{\sum_{i=0}^{N}\hat{Y}_i} = \frac{\sum_{i=0}^{N}(e_i)}{\sum_{i=0}^{N}\hat{Y}_i} \qquad (4)$$

This implies that the best double-talk detection is the echo-canceller itself. Even if the dialog is such that we would stop the system prompt after double-talk was detected, the echo canceller could be needed as a double-talk detector. Based on experimental recognition tests, the double-talk decision performed best when $e_r \geq 0.4$ was the condition for double-talk and for freezing the adaptation.

Comparing figures 3, 4, and 7, we can see the effect of double talk on the relative error function e_r.

5. COMBINING ECHO CANCELLATION AND NOISE REDUCTION

In real-life environments where both additive noise and echo are present (e.g. mobile telephony), the noise reduction module and the echo-cancellation module should take place in obtaining the environment independent features.

Our idea is to first reduce the amount of noise influence (spectral subtraction) and thus enable a better true echo tracing of the echo canceller.

Therefore, we first subtract estimated noise spectra (it can be estimated when neither user speech or

system prompt are present) and then do the subband echo cancelling. In the spectral subtraction of noise, there must be no overestimation noise factor.

This scheme differs greatly from the current post-filtering techniques for noise and residual echo reduction [6], where echo cancellation is performed on noisy input speech.

6. EXPERIMENTAL RESULTS

Recognition experiments were conducted using the following baseline recognizer: Isolated words were recognized using CDHMMs with Laplacian density functions. For each frame (10 ms spaced) we extracted 52 element feature vector. It consisted of mel-scaled cepstral, delta cepstral, end energy components. We used context-dependent diphone models. Each phoneme consisted of 3 segments.

The training and test databases were "Voice Mail" databases (digits and 6 commands) recorded over the German PSTN (analog and digital). Test and training sets are non-overlapping. Results for word accuracy on the clean speech database were 94.9 %.

Two types of echo signals were added to the database: a d1 echo recorded using a digital phone inside the building (signal-to-echo ratio 9.5 dB) and d2 echo over a long distance conversation using a digital telephone (signal-to-echo ratio 4.5 dB). Results are summarized in tables below.

6.1. FFT Approach vs. Reduced Subband Approach

method	echo type	
	d1	d2
No EC performed	76.0 %	22.5 %
FFT subband EC	89.7 %	89.2 %
reduced subband EC	93.3 %	91.3 %

Table 1: Recognition results for system prompt "Die Mailbox des Teilnehmers ist voll...".

6.2. Prompt Embedded Vocabulary

In speech dialog systems, the system and user vocabulary words often overlap. This represents another undesired effect (compare results for no echo cancellation in the table below). However, using the above mentioned double-talk detector and the reduced-subband echo cancelling scheme, the recognition results are greatly improved.

6.3. Combining Echo Cancellation and Noise Reduction

The noise used for combined noise reduction and echo cancellation tests was a car noise at SNR of -5 dB.

method	d1 echo	d1 echo+noise
No EC performed	76.0 %	–
No EC/NR performed	–	13.1 %
EC and NR performed	–	83.2 %

Table 2: Recognition results for a dialog where both echo and car noise are present.

7. CONCLUSIONS

Results infer an absolute necessity of including the echo canceller in a talk-through telephone dialog system. Recognition accuracy was further improved using the proposed reduced-subband approach which in addition to reduced computational load also provides better convergence characteristics of a NLMS algorithm because of increased stopband attenuation of subband division filters (there is less aliasing among subbands after the band reduction). Experiments also show that it is possible to combine echo cancellation and noise reduction by first subtracting estimated noise spectra and then perform the frequency subband echo cancellation.

8. REFERENCES

[1] E. Hänsler. The hands-free telephone problem – an annotated bibliography. *Signal Processing*, 27:259 – 271, 1992.

[2] S. Furui. Adaptive echo cancellation for speech signals. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 597 – 622. Marcel Dekker, Inc., New York, 1992.

[3] A. Gilloire. Adaptive filtering in subbands. In *ICCASP*, pages 2572 – 2575, New York, 1988.

[4] N. Kalouptsidis and S. Theodoridis. *Adaptive System Identification And Signal Processing Algorithms*. Prentice Hall, New York, 1993.

[5] W. Kellermann. *Zur Nachbildung physikalischer Systeme durch parallelisierte digitale Ersatzsysteme im Hinblick auf die Kompensation akustischer Echos*. VDI-Verlag, Düsseldorf, 1989.

[6] R. Martin. Combined acoustic echo cancellation, spectral echo shaping, and noise reduction. In *Proceed. Fourth International Workshop on Acoustic Echo and Noise Control*, pages 48 – 51, Roros, Norway, June 1995.

[7] H. Nishi and M. Kitai. Analysis and detection of double-talk in telephone dialogs. In *Proceed. ICSLP*, pages 111 – 115, Yokohama, 1994.