Adaptive Filter Structures for Enhancing Cocktail Party Speech from Multiple Microphone Recordings

Dirk Van Compernolle [1], *PhD*

Department of Electrical Engineering - ESAT
Katholieke Universiteit Leuven
Kardinaal Mercierlaan 94
3030 Heverlee - Belgium

RESUME

Pour améliorer la qualité de signaux de parole, on propose un filtre adaptative efficace, avec structure Griffiths-Jim. Les différentes sections du filtre sont adaptées indépendentes selon un critère de détection du signal désireé. Les délais sont dérivés à base de computation de corrélations et pour les filtres adaptatives on emploie un algorithme du type LMS. Des expériments avec quatre micros ont étés conduits dans un local avec acoustic réverberant, dans lesquelles on a obtenu une amélioration de 6 à 10 dB SNR.

SUMMARY

In this paper we propose an adaptive filter structure which can significantly enhance cocktail party speech. The underlying structure is a modified "Griffiths-Jim" beamformer in which the sections are selectively adapted on the basis of a signal detection algorithm. The signal detection algorithm relies heavily on the "burst" nature of speech. The look direction is derived from crosschannel correlations and the adaptive filters are adjusted with a standard LMS procedure. The algorithms were tested in a reverberant room with a 4 microphone array. With a faraway talker and non-directional noise typical SNR improvements were 6 to 10dB. Detailed knowledge of the geometry of the array is not required as the algorithms do not rely on perfect spacing of the sensors.

INTRODUCTION

The good performance of human hearing in "cocktail party" situations is helped a lot by the binaural nature of our hearing. By crosscorrelating the input signals we manage to localize and focus on a particular source of interest, especially so for broadband sources. A similar kind of signal enhancement, starting from multiple microphone recordings, will be required if we want speech recognition systems to be successful in the demanding circumstances where several competing sources are present. Currently most speech recognition systems require a SNR of 15dB or better, while many practical applications require an operating range from 0 to +20dB. Speech enhancement on the basis of multimicrophone processing has apart from speech recognition, also applications in transmission systems and in hearing aids [1].

Beamforming for speech applications is complicated by several matters: most rooms are highly reverberant, the speech signal is broadband and the target signal to noise ratio is quite high. Beamformers that rely on minimizing the global output energy fail as soon the SNR is above 0dB and can therefore not be used in a standard speech environment. In this work we use the structure of the Griffiths-Jim beamformer, with an a priori undefined look direction. It was chosen because it allows for separation of determining the look direction and adjusting the adaptive filters for noise cancellation. We found it necessary to achieve a good separation of these problems if we wanted any kind of beamformer to work. Adjustment of look direction and noise cancelling filters will be done under strictly different conditions; the first only is done when the target signal is detected, the second one when the target signal is surely absent. Hence the one assumption on which this work relies is that presence of the target signal can be detected one way or another. This might be on the basis of a moderate SNR or a priori location information.

SETUP

A set of omnidirectional microphones (in our case 4) are placed at respective distances of about 20 cm, a distance roughly equal to the spacing of the human ears. The exact configuration, nor the exact spacing, have much influence on the development of the algorithms. With sound traveling at 340 m/sec and the microphones roughly 20cm apart, we know that the largest possible delay between coherent non-reflected signals is about 0.6 msec. The incoming signals are typically sampled at 20kHz. With good omnidirectional microphones and a single source at a moderate distance from the microphones, the recorded signals will be delayed replicas of each other and scaled by a small gain factor. This gain factor can, apart from directional sensitivity, also reflect small differences in settings of the microphones and amplifiers.

[1] Research Associate of the National Fund for Scientific Research of Belgium (N.F.W.O.)

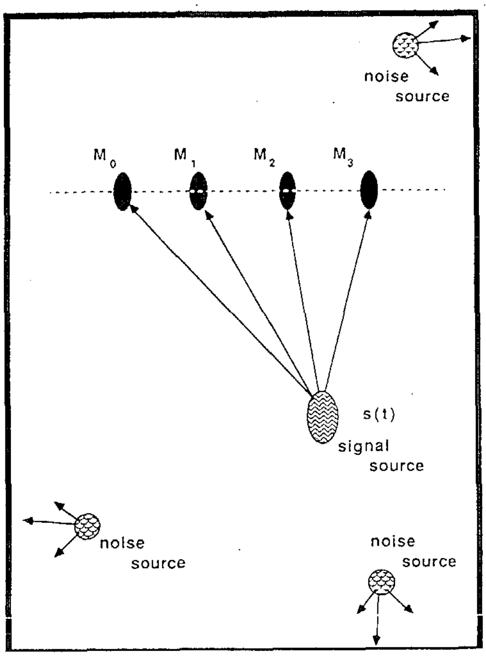

Figure 1: 4-microphone reception of a direct signal and reverberant interference

In a first derivation we will assume that the setup is such that the direct path of the signal $s(t)$ contributes much more to the recordings than the reverberant paths. The interfering noise source $n(t)$ is assumed to be further away from the recording setup and hence can contribute in a very "reverberant" way (Fig. 1). With proper positioning of the microphones these can be realistic assumptions for operating modes of speech recognition systems in a car, a factory floor and for hearing aids. We can write the recorded signals as:

$$y_0(t) = s(t) + n(t) + u_0(t) \quad (1.a)$$
$$y_k(t) = g_k.s(t+\tau_k) + h_k(t) * n(t) + u_k(t) \quad (1.b)$$

in which $n(t)$ represents the correlated part of the noise recordings and $u_k(t)$ the uncorrelated part. The goal of this work is to find an estimate $\hat{s}(t)$ of the signal $s(t)$ as a sum of filtered versions of $y_k(t)$ which has a significantly better signal to noise ratio than either of the recorded signals.

AN ADAPTIVE BEAMFORMER

Structure A multichannel Griffiths-Jim(G-J)[2] beamformer consists of two sections(Fig.2). The first one phase aligns all incoming channels with simple delays. This assumes that the direct path contributions are dominant in the different measurements, which might not always be the case. Sum and differences are computed from all phase aligned channels as indicated in Fig. 2[2] The sum channel $\hat{s_1}(t)$ will serve as signal reference channel:

$$\hat{s}_1(t) = \frac{1}{4}\sum_k \hat{g_k}^{-1} y_k(t - \hat{\tau_k}) \quad (2)$$

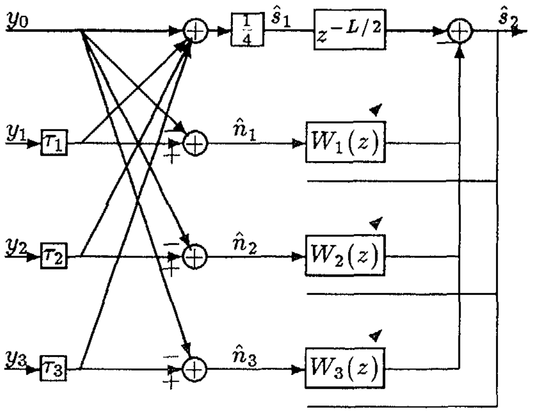

Figure 2: 4-channel Beamformer

The difference channels serve as distinct noise references:

$$\hat{n}_k(t) = \hat{g_k}^{-1} y_k(t - \hat{\tau_k}) - y_0(t) \quad (3)$$

The signal reference and the different noise references are then fed to a classical multichannel adaptive noise canceller ([3],[4]), which yields as final signal estimate:

$$\hat{s}_2(t) = \hat{s}_1(t) - \sum_k w_k(t) * \hat{n}_k(t) \quad (4)$$

This structure has a major potential problem. The noise reference channels will always contain some leakage signal and with a considerable input SNR the leakage component might be quite important. Therefore it is imperative that the filter coefficients $w_k(l)$ are only adapted when no signal at all is present. Furthermore the computation of phasing delays τ_k should be based solely on segments when the target signal was present.

Determining "signal present" condition An advantage of speech is that it occurs in bursts and leaves plenty of time in between the utterances to sample the environment. Conversational speech rarely has a higher than 50% "on"-time. The operation of the beamformer requires the detection of two basic conditions: "signal present" and "signal absent". Therefore the incoming signal is analyzed for several features, among which energy and zerocrossings. Histograms over the most recent past are collected of all features. A multi-modal decomposition of each feature histogram is computed using the EM algorithm. It has been shown[5] that the underlying parameters, and a good decision threshold, can be reliably estimated, although the normal assumption is far from perfect. An example of this, illustrating its effectiveness in poor SNR, is shown in Fig.3. As the region around the threshold still holds some confusion, we used a narrow "no-decision" region in which we classified an incoming frame neither as noise nor speech. The use of frame based parameters (20 msec wide) implies a decision delay, however the use of small no-decision areas around the thresholds allows for using the decision made on the previous frame without noticeable effect.

[2] the gain factors g_k are not indicated in the figure

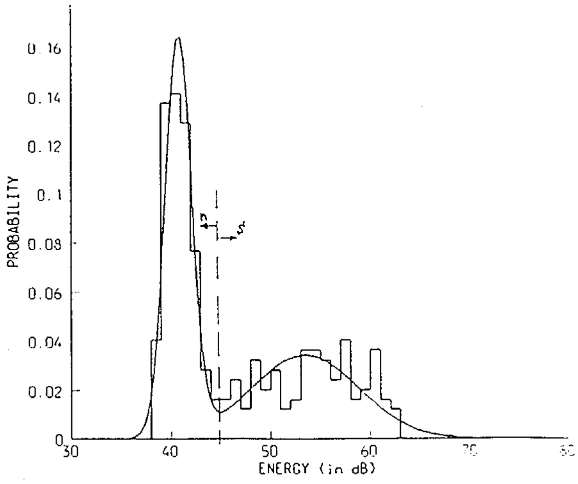

Figure 3: Bimodal Energy Histogram of Noise & Speech

Phase Alignment For the wideband application that speech is, we rely on finding the maximum in the crosscorrelation of the wideband signal between the two channels as a function of delay. A positive signal to noise ratio is required, otherwise the noise might dominate the correlation estimates, yielding senseless results.

Defining $R_{ij}(\tau)$ as :

$$R_{ij}(\tau) = E[y_i(k).y_j(k-\tau)] \tag{5}$$

we derive τ_k and g_k from :

$$\hat{\tau}_k = \max_{\tau} R_{0k}(\tau) \tag{6.a}$$

$$\hat{g}_k^2 = \frac{R_{kk}(0)}{R_{00}(0)} \tag{6.b}$$

We have not attempted interpolation in between consecutive τ-values, hence the resolution of the delay estimate is a single sampling interval. With a sampling rate of 20kHz this corresponds to a delay accuracy of .05 msec or a phase angle of 8 degrees. Higher accuracies or not desirable because of head movement effects during speaking. The expectations $E[]$ in 5) are strictly computed over sections when signal is present.

The intermediate solution $\hat{s}_1(t)$ is the best we can get if all noise sources are uncorrelated (only contributions from $u_k(t)$ in (1)). The maximum theoretical gain in signal to noise ratio for equally loud uncorrelated gaussian noises is only a modest $10log_{10}(k)$ for k microphones. In practical situations it is likely to be no more than 1-3 dB. This is not much, but always useful.

Noise Cancellation A standard LMS algorithm[3] is used for adapting the filters $W_k(z)$ of the second section. The adaptation speed is chosen to be moderate. We want the adaptive filters to reflect the steady interchannel crosscorrelation rather than fast spectral changes in the noise, as in the envisioned applications we do expect nonstationary and slowly moving interfering sources. In reality we achieved adaptation time constants on the order of 100-200 msecs.

DISCUSSION

G-J versus a conventional Noise Canceller We found the G-J structure the most practical one to impose a constraint in the look direction for speech enhancement applications. If the noise is stationary and very much omnidirectional then an unconstrained noise canceller with only adaptation of the coefficients during the signal absent condition would be simpler and also be able to deal with fast moving targets. These conditions however, we do consider as unlikely, and therefore chose the G-J structure.

Problems of Signal Cancellation. In the Griffiths-Jim beamformer the problem of signal cancellation is minimal, provided that the noise vs. speech detection algorithm works properly. If one relies on the beamforming stage to provide a pure noise reference and updates the adaptive filter coefficients at every moment then significant signal cancellation does occur and the adaptive filter alternates between signal cancellation and noise cancellation. The selective updating of the filter coefficients is imperative for successful use of this scheme.

Limitations The connection of the two sections of a G-J beamformer brings along some problems for moving targets. Changing the delays τ_k implies complex adjustments of all filter coefficients $w_k(l)$. This adjustment can furthermore not be done exact. The result is semi-optimal performance of the noise cancelling filters for a short while after modifications to delays. For slowly moving targets this does not pose any significant problems. For fast moving targets the errors can build up to unacceptable limits. It is one of the limitations of this structure for which we have not yet found an adequate solution.

Another drawback of the Griffiths-Jim beamformer is that in cases of very directional noise contributions the simple correlation structure of the noise is distorted in the beamforming stage and cannot be as efficiently exploited as in a simple noise canceller. Overall the robustness given by this scheme seems to far outweigh the suboptimal performance in some circumstances.

EXPERIMENTS

The previous algorithms were implemented and tested for reliability and in informal listening tests. Tests were conducted in two environments: a large office with few reflecting objects and a very reverberant laboratory with a floor space of 3x4m. The array configuration was square or linear with microphone spacings of roughly 20cm. The exact configuration did not have much influence on the obtained results except for some carefully chosen positions of the sources. The speaker was in all experiments about 1m from the microphone. An interfering radio was installed at a similar distance but pointed away from the microphones to give little direct and much reflected sound. It was either badly tuned to provide a "noisy" background or tuned to a station to give a competing speaker or music.

Phasing delays could be accurately estimated in all environments with positive SNRs. In the reverberant space averaging over intervals of about 0.5sec was required so

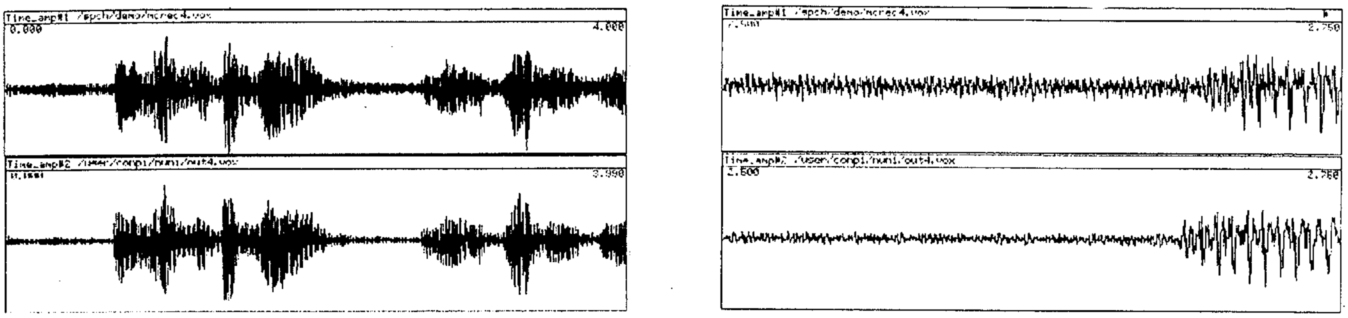

Figure 4: Input and Filtered Signals from Adaptive Beamformer

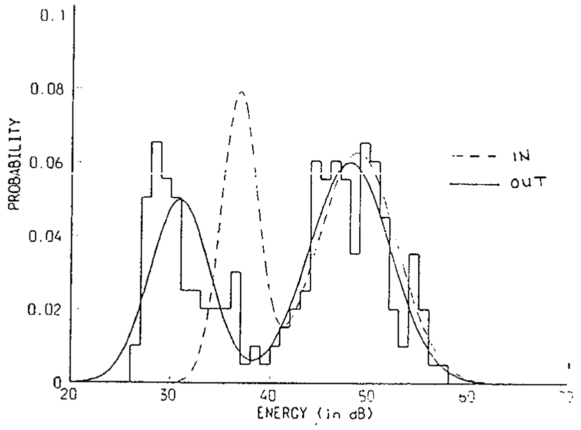

Figure 5: Input and Output Energy Histograms with corresponding Gaussian Decomposition

that only slowly moving sources could be accurately located. The type of background noise did not influence these estimates.

The signal detection algorithm works very fine with the badly tuned radio because the noise energy has a sufficiently narrow variance. The competing speaker case is much more difficult because of the large energy variance of the noise source and requires a larger SNR. Generally speaking our signal detection algorithm needs a slightly higher SNR than the variance of the interfering source. A combination of the single channel based signal detection algorithm with the results of the directionality estimate is possible when we have the a priori knowledge that the target source does not move on a time scale of 1 sec or less. This could significantly enlarge the SNR operating range for the competing speaker case.

Fig. 4 shows the input and output time waveforms in the case of the badly tuned radio. Fig. 5 illustrates the corresponding input and output energy histograms and their gaussian fits. An overall gain of about 7dB can be seen. The improvement between input and output is very clearly audible with no distortion whatsoever. This result is typical for many tested conditions. It is less than what theoretically could be predicted and what has been found in many simulation experiments. However we consider it quite successful given a very reverberant room and a fluctuating interference.

CONCLUSIONS

A multimicrophone noise suppression algorithm was successfully implemented with possible applications in speech enhancement, speech recognition or hearing aids. A key feature is the accurate detection of moments when the target signal is present. The algorithms used in this work are intended for input SNR conditions of 5 to 20dB and work in a stationary background as well in the case of a competing speaker. In the competing speaker case it is significantly more difficult to decide when the target signal is on. A priori knowledge about the moving dynamics of the target is very helpful in this case. More complex decision rules which use all possible information about the target of interest will allow us to improve the accuracy of the target detection and consequently extend the operating range of this type of speech enhancement system.

References

[1] **D. Van Compernolle** . Hearing Aids using Binaural Processing Principles. In *International Symposium on Digital Processing Hearing Aids, Wolfheze NL*, Aug. 1988. Also to be published in Acta-Otolaryngologica.

[2] **L.J. Griffiths and C.W. Jim** . An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.*, vol. AP-30:27–34, Jan. 1982.

[3] **B.Widrow and S.D.Stearns**. *Adaptive Signal Processing.* Prentice-Hall, 1985.

[4] **E. Ferrara and B. Widrow** . Multichannel Adaptive Filtering for Signal Enhancement . *IEEE Trans. Acoust., Speech, Signal Processing* , vol. ASSP 29.3:766–770, 1981.

[5] **D. Van Compernolle**. Noise adaptation in a hidden markov model speech recognition system. *Computer Speech and Language*, 1989. To be published.